Hypothesis_Testing

Tejovardhan Medamarti

February 8, 2017

Overview and summary

Test hypotheses for the price of automobiles:

Source of the data can be found at : https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data

Our objective is to find the hypothesis for the given below questions.

1. Compare and test Normality the distributions of price and log price - Use both a graphical method and a formal test.
2. Test significance of price (log price) stratified by a) fuel type, b) aspiration, and c) rear vs. front wheel drive. Use both graphical methods and the formal test.
3. Apply ANOVA to the auto price data to compare the price (or log price if closer to a Normal distribution) of autos stratified by number of doors, and body style - two sets of tests.

- Graphically explore the differences between the price conditioned by the categories of each variable -?Hint, make sure you have enough data for each category.
- Use standard ANOVA and Tukey ANOVA to test the differences of these groups.

Note: Following packages are required to run the below report.

- dplyr

```
rm(list = ls())
require(dplyr)
require(ggplot2)

setwd("C:\\Tejo\\DataScience\\UW_Datascience_Course\\350\\DataScience350-master\\Lecture4\\Assignment")
```

Data loading and preparation

```
read.auto = function(file = 'Automobile price data _Raw_.csv'){
  ## Read the csv file
  auto.price <- read.csv(file, header = TRUE,
                         stringsAsFactors = FALSE)

  ## Coerce some character columns to numeric
  numcols <- c('price', 'bore', 'stroke', 'horsepower', 'peak.rpm')
  auto.price[, numcols] <- lapply(auto.price[, numcols], as.numeric)

  ## Remove cases or rows with missing values. In this case we keep the
  ## rows which do not have nas.
  auto.price[complete.cases(auto.price), ]
}
auto.price = read.auto()
```

```
## Warning in lapply(auto.price[, numcols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(auto.price[, numcols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(auto.price[, numcols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(auto.price[, numcols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(auto.price[, numcols], as.numeric): NAs introduced by
## coercion
```

```
#Compare and test Normality the distributions of price and log price
#- Use both a graphical method and a formal test.

read.auto <- auto.price[auto.price$drive.wheels != "4wd",c("fuel.type",
                                        "aspiration","drive.wheels" ,"price" ,
                                        "num.of.doors", "body.style" )]

read.auto <- read.auto[read.auto$num.of.doors != "?", ]
read.auto$log.price <- log(read.auto$price)
read.auto$scaled.log.price <- scale(read.auto$log.price, center = TRUE, scale = TRUE)
read.auto$scaled.price <- scale(read.auto$price, center = TRUE, scale = TRUE)

pop_auto.price = rnorm(nrow(read.auto), mean=mean(read.auto$scaled.price), sd = sd(read.auto$scaled.pri
pop_auto.norm = rnorm(nrow(read.auto), mean=0, sd = 1)
pop_auto.log.price = rnorm(nrow(read.auto), mean=mean(read.auto$scaled.log.price), sd = sd(read.auto$sca
```

Lets start with the Question 1:

Compare and test Normality the distributions of price and log price - Use both a graphical method and a formal test.

Null Hypothesis: Distribution of log price data is identical to Standard Normal distribution

```
ks.test(read.auto$scaled.log.price, pop_auto.norm)
```

```
## Warning in ks.test(read.auto$scaled.log.price, pop_auto.norm): p-value will
## be approximate in the presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  read.auto$scaled.log.price and pop_auto.norm
## D = 0.10811, p-value = 0.2298
## alternative hypothesis: two-sided
```

- Based on the above results, we **ACCEPT** the null hypothesis as the P-values is > 0.05.

Null Hypothesis: Distribution of price data is identical to Standard Normal distribution

```
ks.test(read.auto$scaled.price, pop_auto.norm)
```

```
## Warning in ks.test(read.auto$scaled.price, pop_auto.norm): p-value will be
## approximate in the presence of ties
```

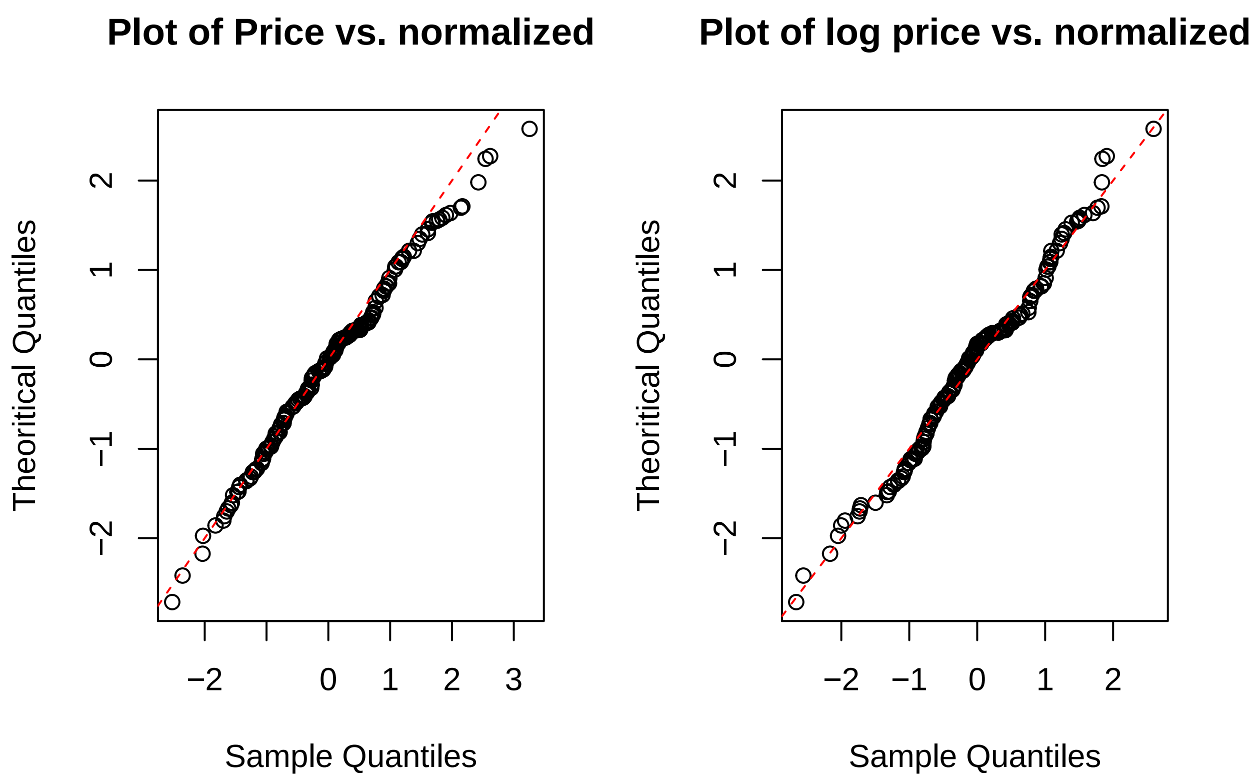
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  read.auto$scaled.price and pop_auto.norm
## D = 0.17838, p-value = 0.005553
## alternative hypothesis: two-sided
```

Conclusion: Based on the above results, we **REJECT** the null hypothesis as the P-value is very small.

Visualization:

Lets draw the QQ plot for the price and log price.

```
par(mfrow = c(1, 2))

plot(sort(pop_auto.price), sort(pop_auto.norm), main = 'Plot of Price vs. normalized',
     xlab = 'Sample Quantiles', ylab = 'Theoritical Quantiles')
abline(a = 0.0, b = 1.0, lty = 2, col = 'red')

plot(sort(pop_auto.log.price), sort(pop_auto.norm), main = 'Plot of log price vs. normalized',
     xlab = 'Sample Quantiles', ylab = 'Theoritical Quantiles')
abline(a = 0.0, b = 1.0, lty = 2, col = 'red')
```



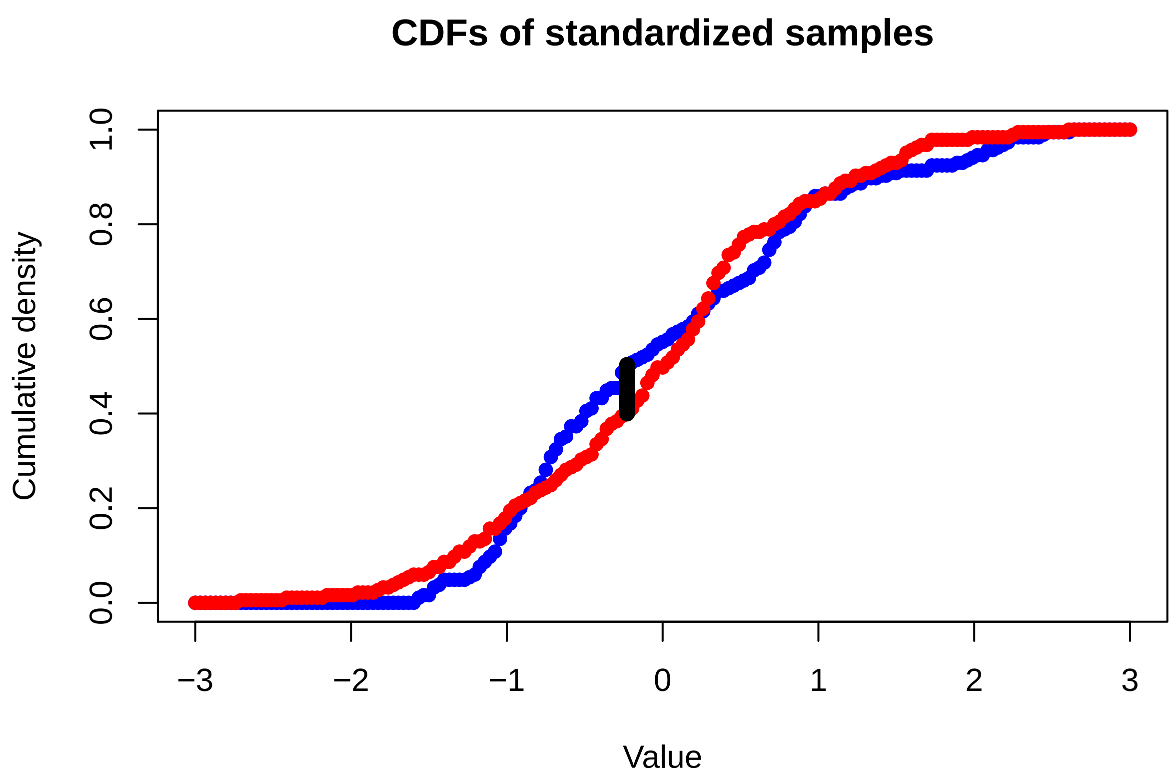

```
par(mfrow = c(1, 1))
```

Lets draw a CDF plot for log price vs standard normalized data.

```r
# Have to standardize the x-values
x_seq = seq(-3,3,len=nrow(read.auto))
y_cdf1 = sapply(x_seq, function(x){
        sum(read.auto$scaled.log.price<x)/length(read.auto$scaled.log.price)
})
y_cdf2 = sapply(x_seq, function(x){
        sum(pop_auto.norm<x)/length(pop_auto.norm)
})

## Find the max deviation
k_s_stat = max(abs(y_cdf1 - y_cdf2))
k_s_stat
```

```
## [1] 0.1027027
```

```r
# where does it occur?
k_index = which.max(abs(y_cdf1-y_cdf2))
k_s_x = x_seq[k_index]
plot(x_seq,y_cdf1, col='blue', pch=16, main ='CDFs of standardized samples',
     xlab = 'Value', ylab = 'Cumulative density')
points(x_seq,y_cdf2,col='red', pch=16)
lines(c(k_s_x,k_s_x), c(y_cdf1[k_index],y_cdf2[k_index]),
      col='black', lwd=8)
```



Lets do our second question:

2. Test significance of price (log price) stratified by a) fuel type, b) aspiration, and c) rear vs. front wheel drive. Use both graphical methods and the formal test.

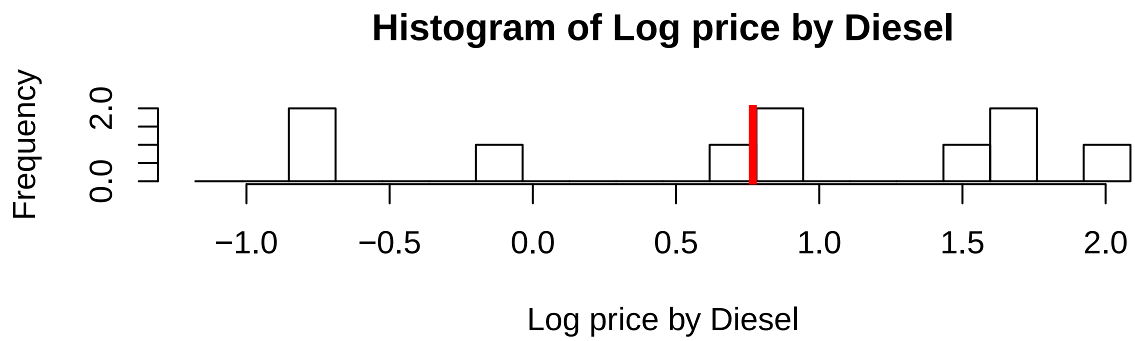
```
plot.t <- function(a, b, cols = c('pop_A', 'pop_B'), nbins = 20){
        maxs = max(c(max(a), max(b)))
        mins = min(c(min(a), min(b)))
        breaks = seq(maxs, mins, length.out = (nbins + 1))
        par(mfrow = c(2, 1))
        hist(a, breaks = breaks, main = paste('Histogram of', cols[1]), xlab = cols[1])
        abline(v = mean(a), lwd = 4, col = 'red')
        hist(b, breaks = breaks, main = paste('Histogram of', cols[2]), xlab = cols[2])
        abline(v = mean(b), lwd = 4, col = 'red')
        par(mfrow = c(1, 1))
}
```

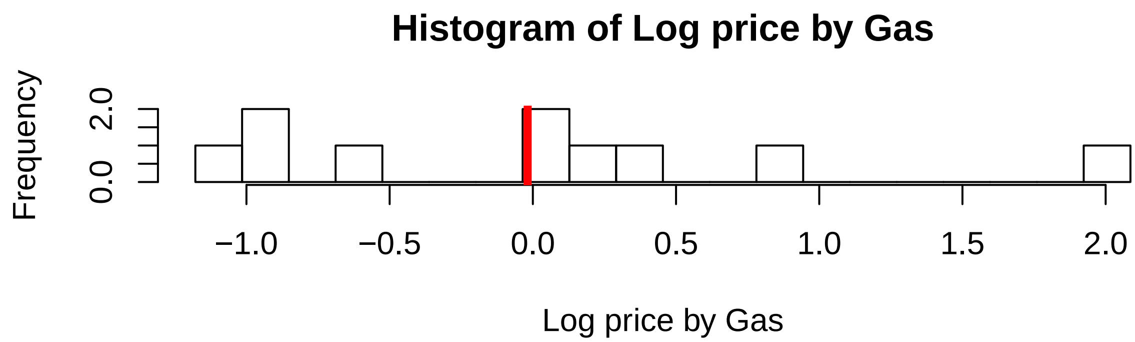
For this analysis, i am stratifying the data by 10 records for each group.

2(a). Lets start the analysis for log price by fuel type.

Null Hypothesis: Significance of log price by fuel type. There is no price difference with the fuel type.

```
n = 10

auto.fuel.type = read.auto %>% group_by(fuel.type ) %>%
        sample_n(n, replace = FALSE)

auto.fuel.type.bydiesel <-  auto.fuel.type[auto.fuel.type$fuel.type=="diesel",]
auto.fuel.type.bygas <-  auto.fuel.type[auto.fuel.type$fuel.type=="gas",]

pop_A = auto.fuel.type.bydiesel$scaled.log.price
pop_B = auto.fuel.type.bygas$scaled.log.price

plot.t(pop_A, pop_B, cols = c("Log price by Diesel", "Log price by Gas"))
```






```
ks.test(pop_A, pop_B, alternative = "two.sided")
```

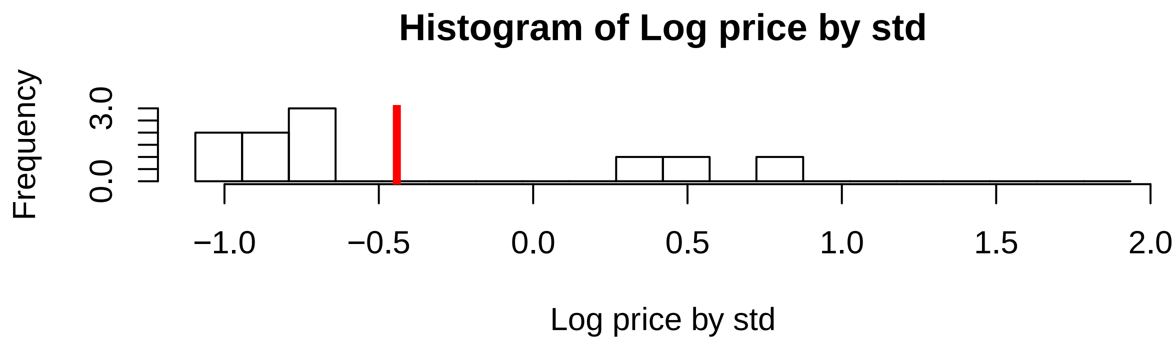
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  pop_A and pop_B
## D = 0.5, p-value = 0.1678
## alternative hypothesis: two-sided
```

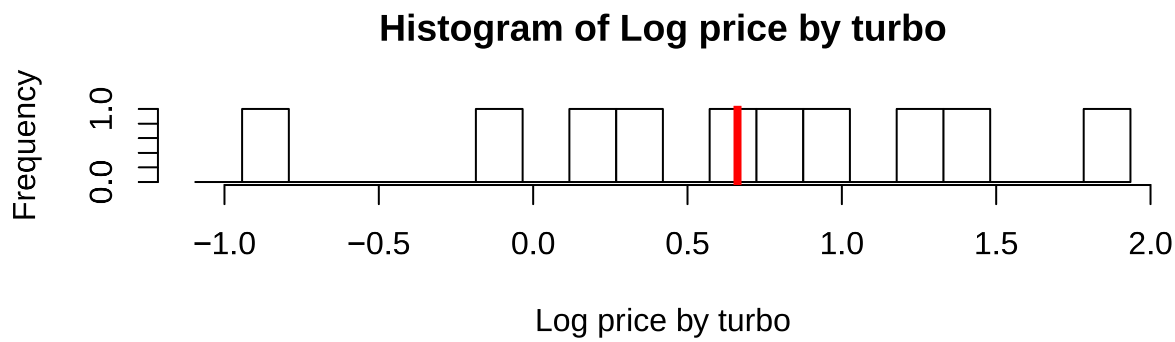
Conclusion: Based on the above results, we failed to reject, hence we **ACCEPT** the null hypothesis.

2(b). Lets start the analysis for log price by aspiration.

Null Hypothesis: Significance of log price by aspiration. There is no price difference with the aspiration.

```
auto.aspiration.type = read.auto %>% group_by(aspiration) %>%
        sample_n(n, replace = FALSE)

auto.aspiration.type.std <-  auto.aspiration.type[auto.aspiration.type$aspiration=="std",]
auto.aspiration.type.turbo <-  auto.aspiration.type[auto.aspiration.type$aspiration=="turbo",]


pop_A = auto.aspiration.type.std$scaled.log.price
pop_B = auto.aspiration.type.turbo$scaled.log.price

plot.t(pop_A, pop_B, cols = c("Log price by std", "Log price by turbo"))
```






```
ks.test(pop_A, pop_B, alternative = "two.sided")
```

```
## Warning in ks.test(pop_A, pop_B, alternative = "two.sided"): cannot compute
## exact p-value with ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  pop_A and pop_B
## D = 0.6, p-value = 0.05465
## alternative hypothesis: two-sided
```

Conclusion: Based on the above results, we **ACCEPT** the null hypothesis.

2(C). Lets start the analysis for log price by drive wheels

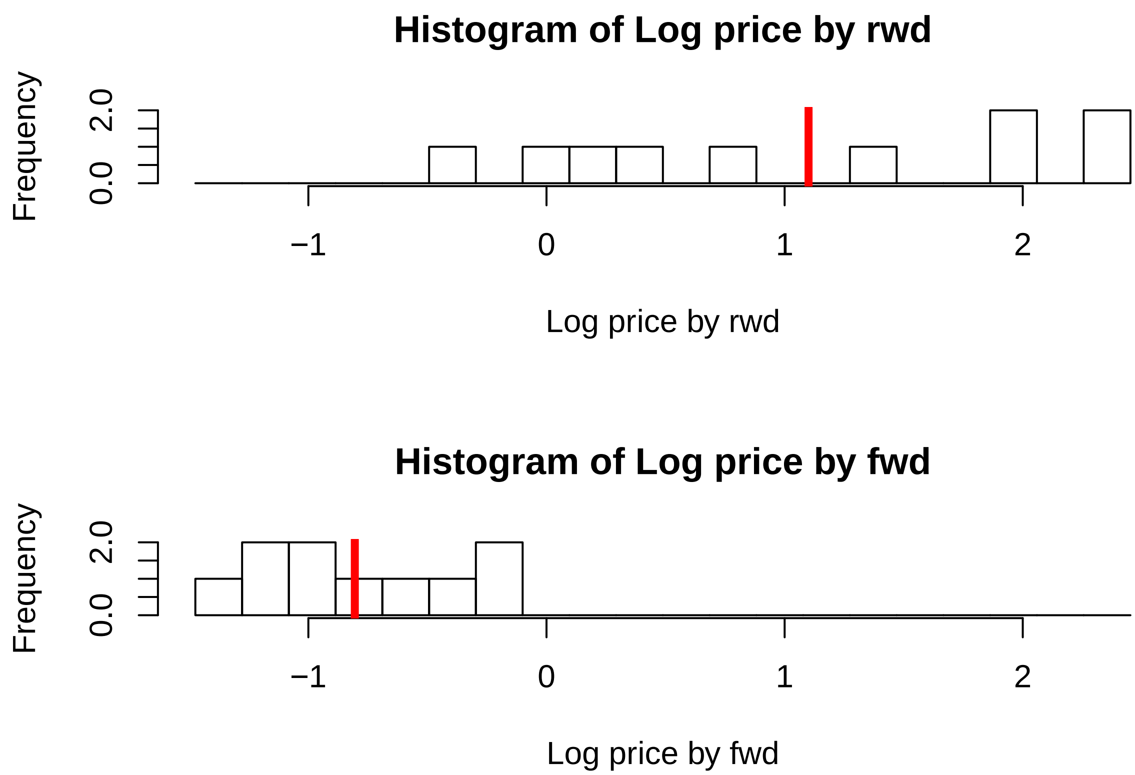
Null Hypothesis: Significance of log price by drive wheels There is no price difference with the drive wheels.

```
auto.drive.wheels = read.auto %>% group_by(drive.wheels) %>%
        sample_n(n, replace = FALSE)

auto.drive.wheels.rwd <-  auto.drive.wheels[auto.drive.wheels$drive.wheels=="rwd",]
auto.drive.wheels.fwd <-  auto.drive.wheels[auto.drive.wheels$drive.wheels=="fwd",]


pop_A = auto.drive.wheels.rwd$scaled.log.price
pop_B = auto.drive.wheels.fwd$scaled.log.price
#plot.t(pop_A, pop_B)
```

```r
ks.test(pop_A, pop_B, alternative = "two.sided")
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  pop_A and pop_B
## D = 0.9, p-value = 0.0002165
## alternative hypothesis: two-sided
```

```r
plot.t(pop_A, pop_B, cols = c("Log price by rwd", "Log price by fwd"))
```




```r
ks.test(pop_A, pop_B, alternative = "two.sided")
```

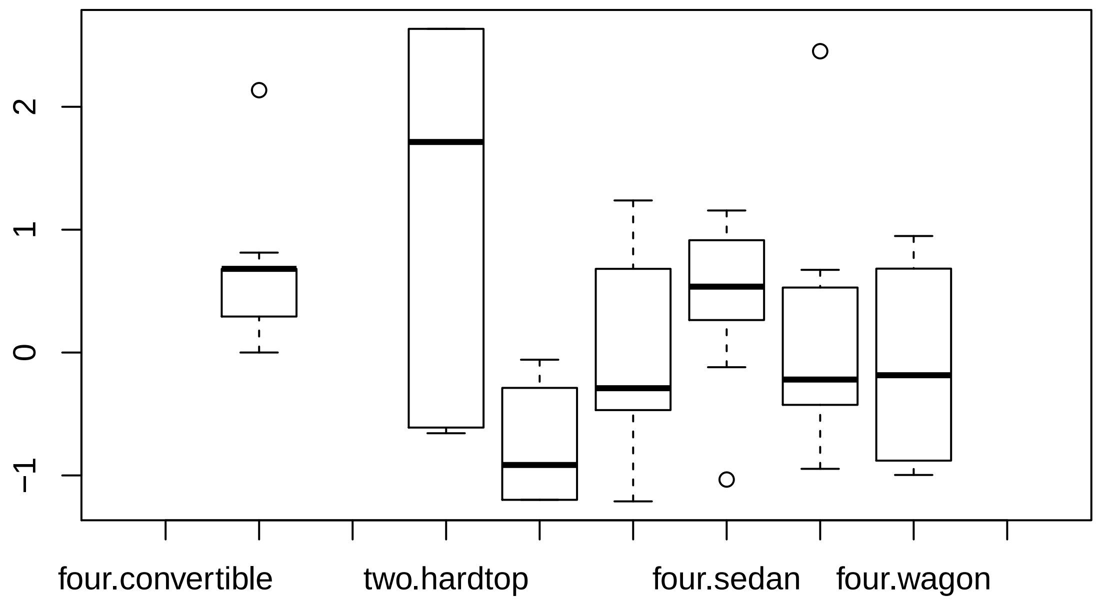
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  pop_A and pop_B
## D = 0.9, p-value = 0.0002165
## alternative hypothesis: two-sided
```

Conclusion: Based on the above results, we **ACCEPT** the null hypothesis.

3. Apply ANOVA to the auto price data to compare the price (or log price if closer to a Normal distribution) of autos stratified by number of doors, and body style - two sets of tests.

Null Hypothesis: Significance of log price by Num of doors and body style. There is no price variances with the influence of Num of doors and body style.

```r
auto.anova.sample = read.auto %>% group_by(num.of.doors, body.style) %>%
        sample_n(n, replace = TRUE)

boxplot(auto.anova.sample$scaled.log.price ~ auto.anova.sample$num.of.doors+
                auto.anova.sample$body.style)
```




```r
df_aov = aov(scaled.log.price ~ num.of.doors + body.style, data = auto.anova.sample)
summary(df_aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## num.of.doors  1   6.15   6.152   7.735 0.00711 **
## body.style    4  13.71   3.428   4.309 0.00378 **
## Residuals    64  50.90   0.795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(df_aov)
```

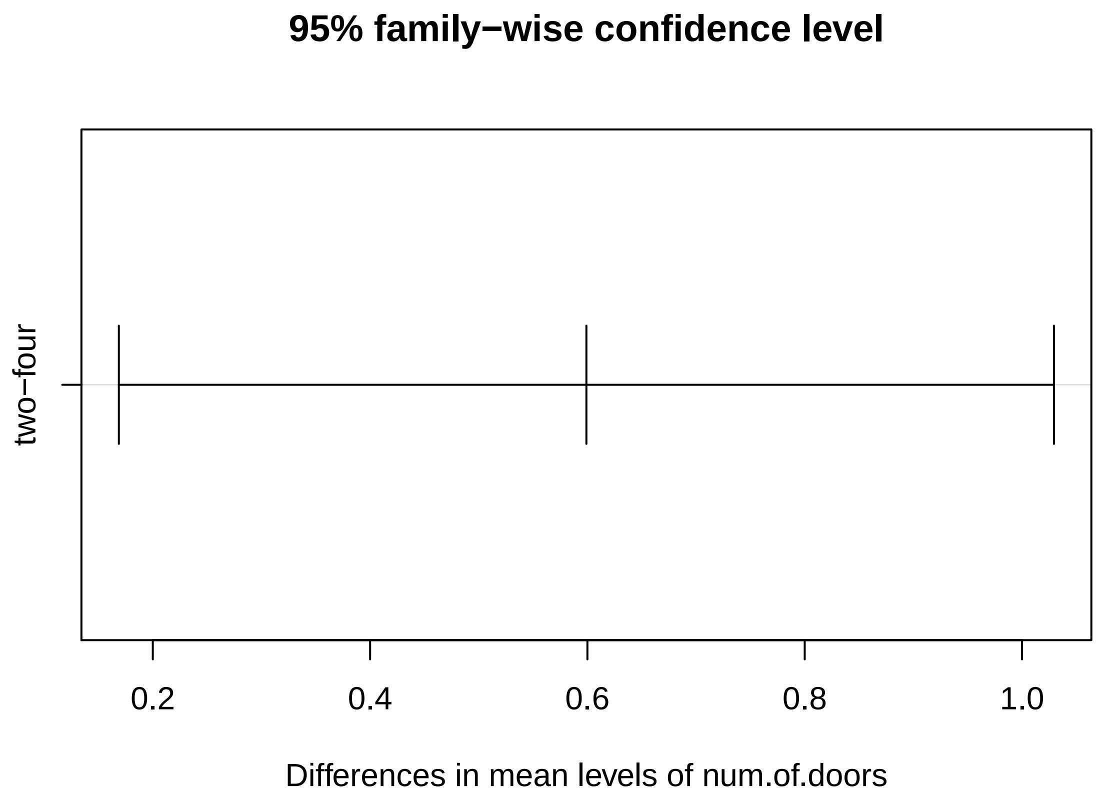
```
## Call:
##    aov(formula = scaled.log.price ~ num.of.doors + body.style, data = auto.anova.sample)
##
## Terms:
##                 num.of.doors body.style Residuals
## Sum of Squares       6.15221   13.71027  50.90402
## Deg. of Freedom            1          4        64
##
## Residual standard error: 0.8918381
```

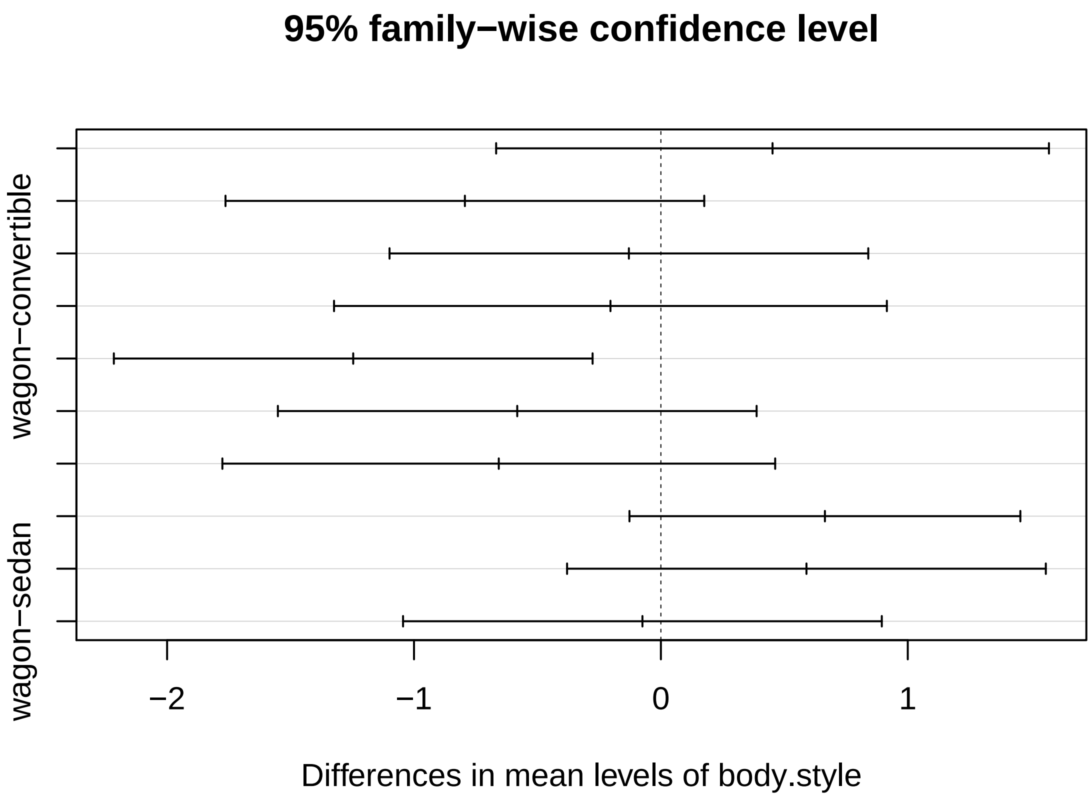
```
## Estimated effects may be unbalanced
```

```r
tukey_anova = TukeyHSD(df_aov)  # Tukey's Range test:
tukey_anova
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = scaled.log.price ~ num.of.doors + body.style, data = auto.anova.sample)
##
## $num.of.doors
##               diff       lwr      upr     p adj
## two-four 0.5990651 0.1687554 1.029375 0.0071066
##
## $body.style
##                               diff        lwr        upr     p adj
## hardtop-convertible    0.45229641 -0.6672801  1.5718729 0.7879444
## hatchback-convertible -0.79390614 -1.7634878  0.1756755 0.1587835
## sedan-convertible     -0.12953503 -1.0991167  0.8400466 0.9956879
## wagon-convertible     -0.20425438 -1.3238308  0.9153221 0.9858664
## hatchback-hardtop     -1.24620255 -2.2157842 -0.2766209 0.0053005
## sedan-hardtop         -0.58183143 -1.5514131  0.3877502 0.4506562
## wagon-hardtop         -0.65655078 -1.7761272  0.4630257 0.4741616
## sedan-hatchback        0.66437112 -0.1272890  1.4560312 0.1411279
## wagon-hatchback        0.58965177 -0.3799299  1.5592334 0.4369648
## wagon-sedan           -0.07471935 -1.0443010  0.8948623 0.9995001
```

```r
plot(tukey_anova)
```


95% family−wise confidence level
two−four
0.2
0.4
0.6
0.8
1.0
Differences in mean levels of num.of.doors



Conclusion: Based on the above results, 1. "hatchback-convertible" body style has significant impact on the log price, hence we are **rejecting** the null hypothesis and for the rest of the other body styles, we are **accepting**.

2. For num.of.doors, we are **rejecting** the null hypothesis.

Conclusion:

Log price is having signifant difference only when, - body style equals "hatch back convertible" - the num of doors.