Ryan Timbrook
Data Science 350 – Homework Assignment 5

Assignment:

Apply bootstrap resampling to the auto price data as follows:

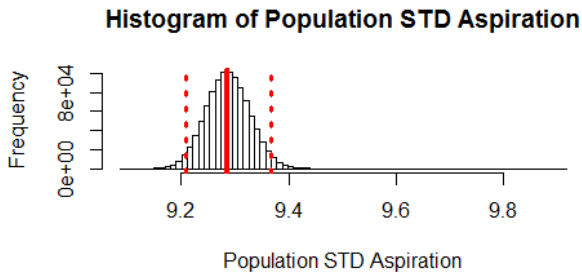
- Compare the difference of the bootstrap resampled mean of the log price of autos grouped by 1) aspiration and 2) fuel type. Use both numerical and graphical methods for your comparison. Are these means different within a 95% confidence interval? How do your conclusions compare to the results you obtained using the t-test last week?
- Compare the differences of the bootstrap resampled mean of the log price of the autos grouped by body style. You will need to do this pair wise; e.g. between each possible pairing of body styles. Use both numerical and graphical methods for your comparison. Which pairs of means are different within a 95% confidence interval? How do your conclusions compare to the results you obtained from the ANOVA and Tukey's HSD analysis you performed last week?

Observations:

- Difference of bootstrap resampled mean of the log price of autos
    - Grouped By:
        - Aspiration:
            - The distribution of the bootstrap means do not overlap. Their difference is significant. We can reject the null hypothesis that std and turbo aspirated cars mean prices are the same. This is represented in Table 1 below.
            - This is consistent with the conclusion produced when using the t-test. This is represented in Table 3 below.
        - Fuel Type:
            - The distribution of the bootstrap means overlap. We cannot reject the null hypothesis at 95% confidence that these means are the same. This is represented in Table 2 below.
            - This is consistent with the conclusion produced when using the t-test. This is represented in Table 4 below.
        - Body Styles:
            - The following pairs of boot strap mean distributions are different based on a 95% confidence level, we reject the null hypothesis for these pairs that their means are the same. This is represented in Tables 7.2, 7.5 and 7.8 below
                - hatchback-convertible
                - hatchback-hardtop
                - sedan-hatchback
            - This is consistent with the conclusion produced when using the ANOVA and Tukey HSD analysis. This is represented in Tables 5 and 6 below.

Table 1: Bootstrap resample of mean std aspiration and turbo aspiration

Observation: The distribution of the bootstrap means do not overlap. Their difference is significant. We can reject the null hypothesis that std and turbo aspirated cars mean prices are the same.



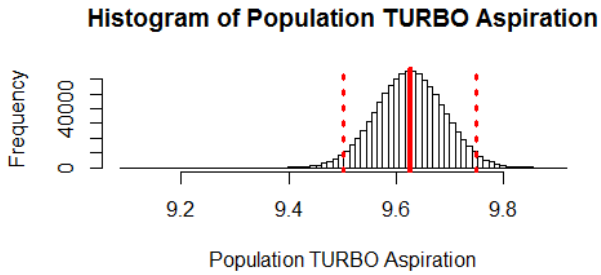


Table 1.1: Bootstrap difference of means std aspiration and turbo aspiration

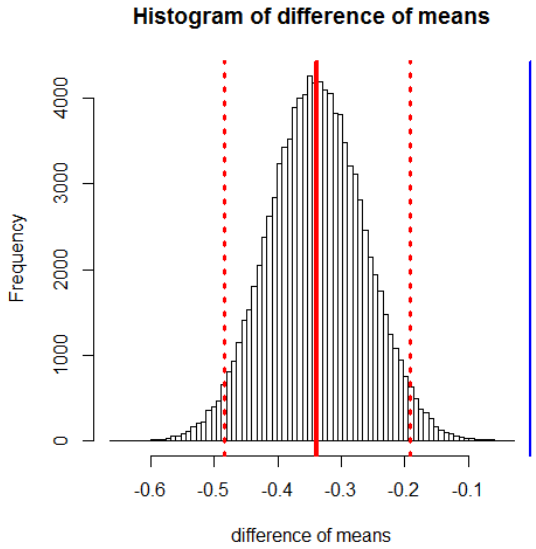


Table 1.2: Q-Q normal plot of bootstrap difference in means

The points on the Q-Q normal plot are nearly on a straight line. The bootstrap difference in means conforms to the CLT.

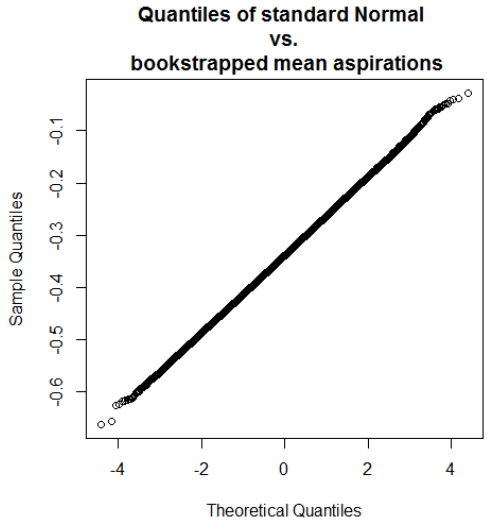
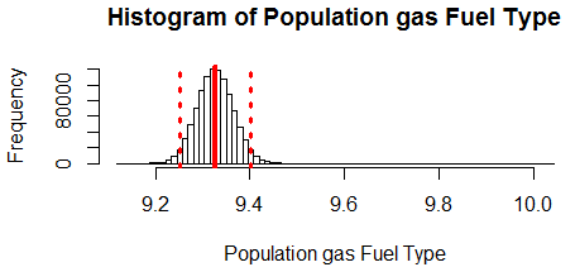


Table 2: Bootstrap resample of mean gas fuel type and diesel fuel type

Observation: The distribution of the bootstrap means overlap. We cannot reject the null hypothesis at 95% confidence that these means are the same.



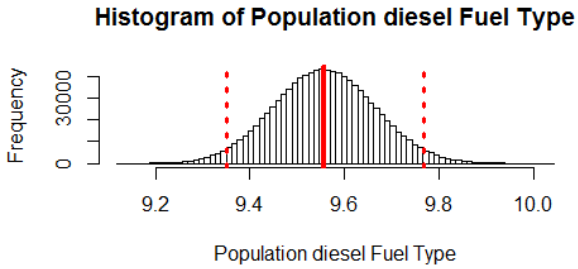


Table 2.1: Bootstrap difference of means gas fuel type and diesel fuel type

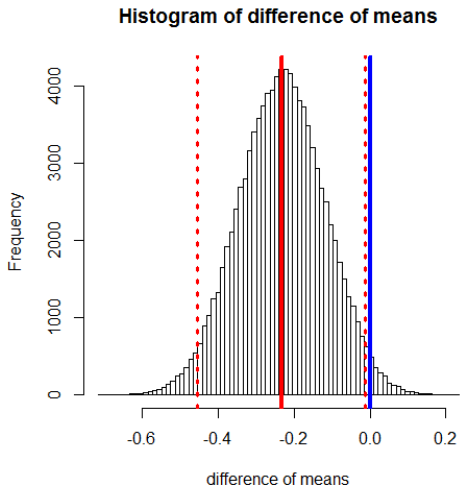


Table 2.2: Q-Q normal plot of bootstrap difference in means

The points on the Q-Q normal plot are nearly on a straight line. The bootstrap difference in means conforms to the CLT.

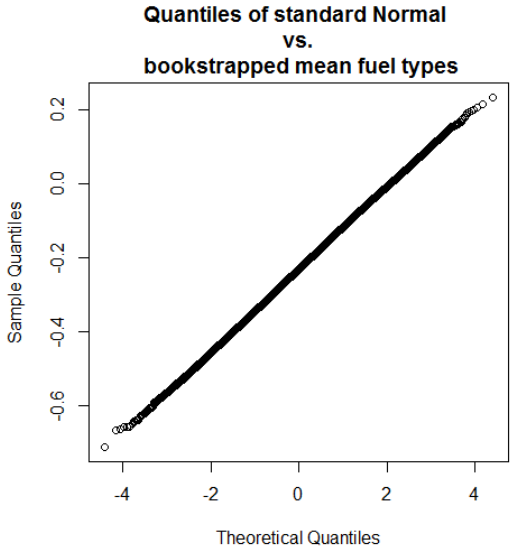


Table 3: Significance test, Price comparison by Diesel vs. Gas Fueled Cars

At 95% confidence we **cannot** reject the null hypothesis that these means are the same. The p-value is greater than .025 and the confidence interval overlaps zero.

```
        Welch Two Sample t-test

data:  diesel.lnprices and gas.lnprices
t = 1.9397, df = 24.363, p-value = 0.06408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01424314  0.46494692
sample estimates:
mean of x mean of y
 9.557420  9.332068
```

	fuel.type	count	mean.price	mean.lnprice	sd.price	sd.lnprice	max.price	max.lnprice	min.price	min.lnprice
1	diesel	20	15838.15	9.557420	7759.844	0.4880124	31600	10.36091	7099	8.867709
2	gas	167	13081.87	9.332068	8199.532	0.5152990	45400	10.72327	5118	8.540519

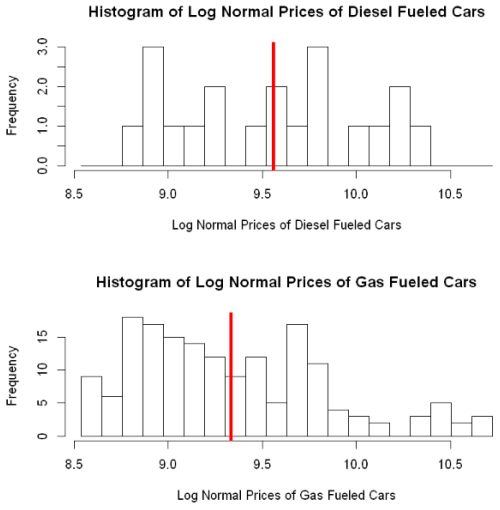


Table 4: Significance test, Price comparison by Turbo vs. Standard Cars

At 95% confidence we **can** reject the null hypothesis that these means are the same. The p-value is significantly less than .025 and the confidence interval does not overlaps zero.

```
        Welch Two Sample t-test

data:  std.lnprices and turbo.lnprices
t = -4.44, df = 62.417, p-value = 3.742e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5071209 -0.1922786
sample estimates:
mean of x mean of y
 9.292588  9.642288
```

	aspiration	count	mean.price	mean.lnprice	sd.price	sd.lnprice	max.price	max.lnprice	min.price	min.lnprice
1	std	153	12674.61	9.292588	8404.835	0.5202030	45400	10.72327	5118	8.540519
2	turbo	34	16535.88	9.642288	6247.721	0.3883027	31600	10.36091	7689	8.947546

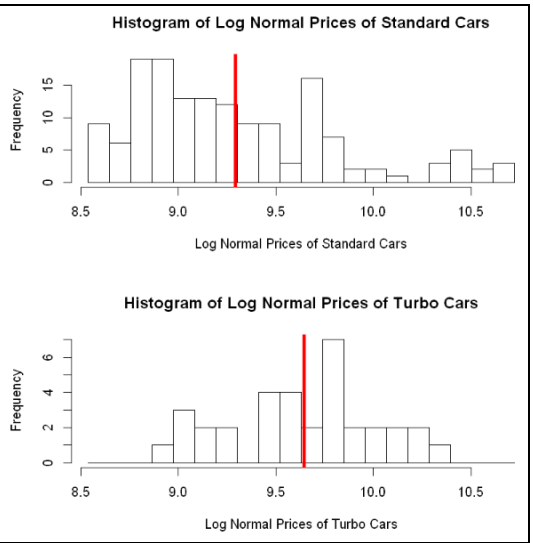
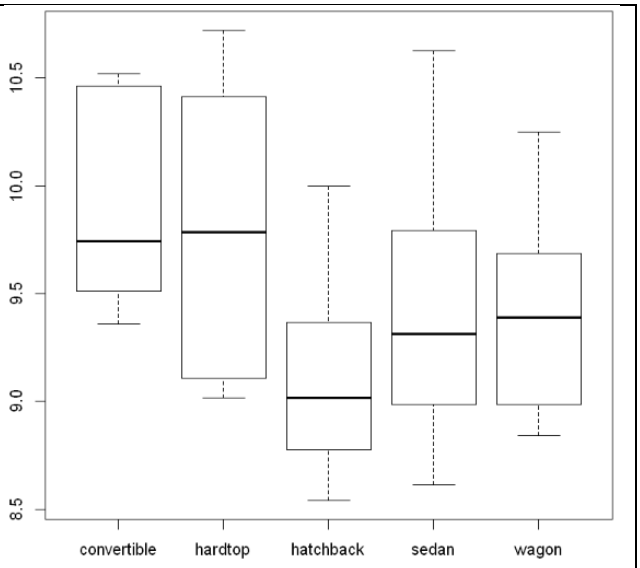


Table 5: Boxplot Graph of LN Auto Prices by Body Style

Body Style **has a significant** impact on auto prices. Based on the high F statistic shown below and the very small p-value we can reject the null hypothesis that these groups mean values are the same for all body styles

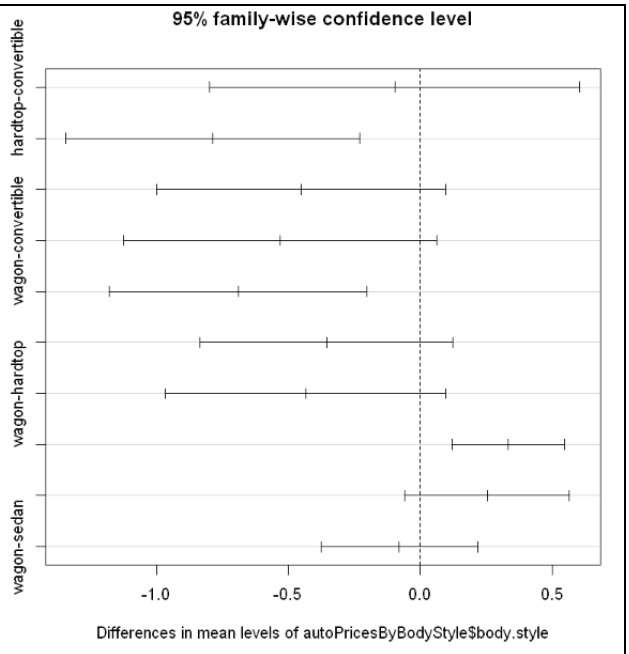


ANOVA Summary Data:

```
                                Df Sum Sq Mean Sq F value   Pr(>F)
autoPricesByBodyStyle$body.style   4   7.85  1.9615   8.788 1.57e-06 ***
Residuals                        190  42.41  0.2232
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
   aov(formula = autoPricesByBodyStyle$lnprice ~ autoPricesByBodyStyle$body.style)

Terms:
                autoPricesByBodyStyle$body.style Residuals
Sum of Squares                           7.84591  42.41013
Deg. of Freedom                                4       190

Residual standard error: 0.4724523
Estimated effects may be unbalanced
```

Table 6: Tukey ANOVA – HSD Test

Body Style **has a significant** impact on auto prices. We can reject the null hypothesis that body style mean prices are the same for all groupings. The graph and data summary below shows seven of the groups cross over the zero line representing a significant difference in mean values.



Summary Data:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = autoPricesByBodyStyle$lnprice ~ autoPricesByBodyStyle$body.style)

$`autoPricesByBodyStyle$body.style`
                      diff         lwr         upr     p adj
hardtop-convertible   -0.09664988 -0.79938112  0.60608136 0.9955964
hatchback-convertible -0.78537118 -1.34130681 -0.22943556 0.0012903
sedan-convertible     -0.45193455 -0.99984087  0.09597177 0.1586910
wagon-convertible     -0.53101926 -1.12493556  0.06289704 0.1037126
hatchback-hardtop     -0.68872130 -1.17710344 -0.20033917 0.0013238
sedan-hardtop         -0.35528467 -0.83450698  0.12393764 0.2502185
wagon-hardtop         -0.43436938 -0.96558426  0.09684551 0.1654127
sedan-hatchback        0.33343663  0.12157052  0.54530274 0.0002276
wagon-hatchback        0.25435193 -0.05777382  0.56647767 0.1680903
wagon-sedan           -0.07908470 -0.37667401  0.21850460 0.9488191
```

Table: 7.1: Bootstrap difference of means, hardtop-convertible

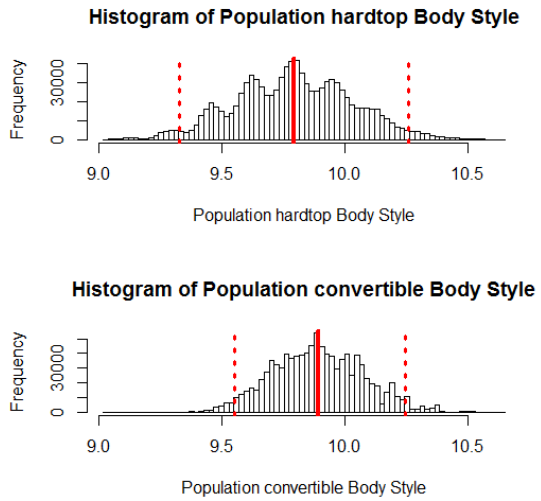


***Table 7.2: Bootstrap difference of means, hatchback-convertible**

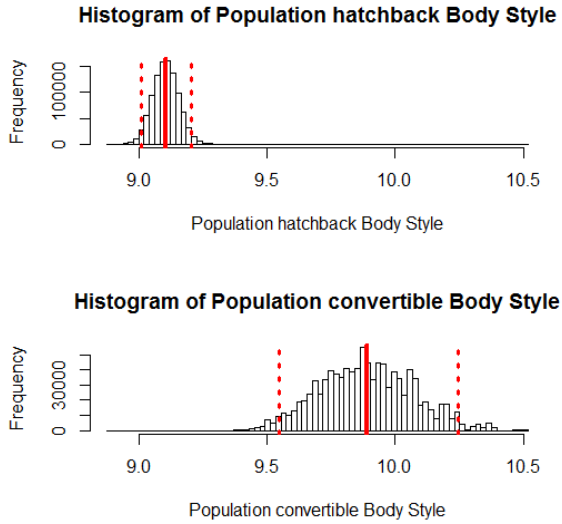


Table 7.3: Bootstrap difference of means, sedan-convertible

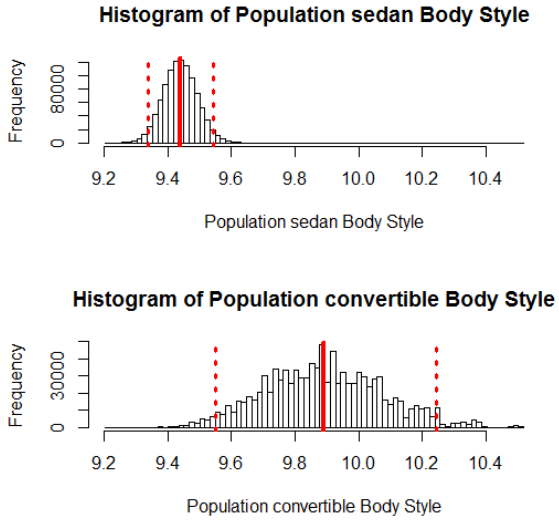


Table 7.4: Bootstrap difference of means, wagon-convertible

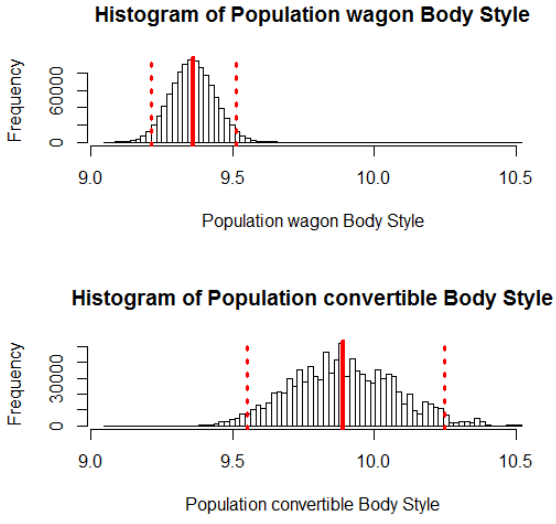
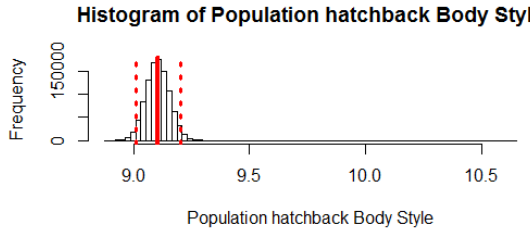


***Table 7.5: Bootstrap difference of means, hatchback-hardtop**



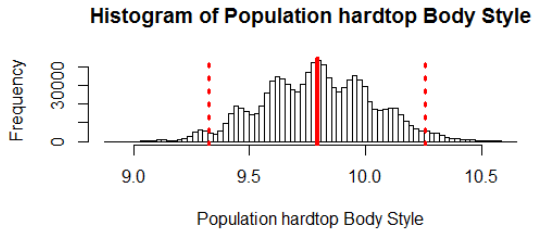
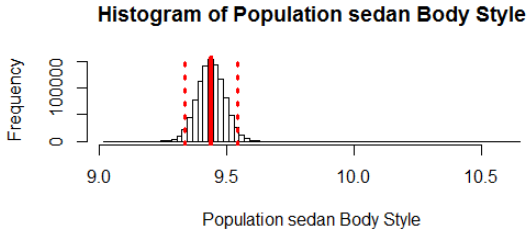


Table 7.6: Bootstrap difference of means, sedan-hardtop



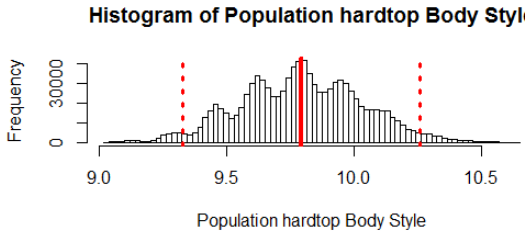
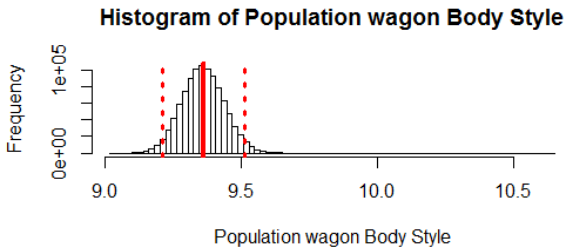


Table 7.7: Bootstrap difference of means, wagon-hardtop



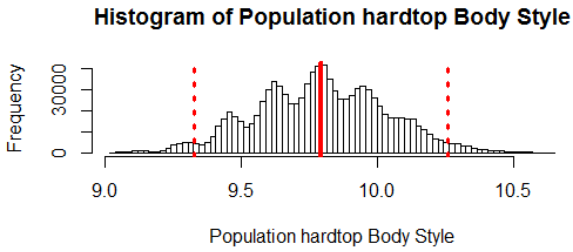
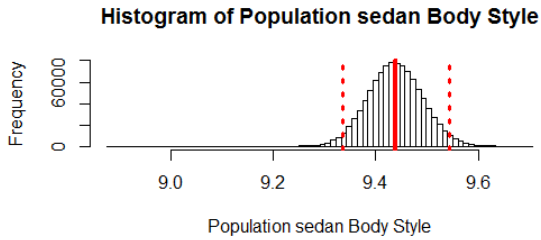


***Table 7.8: Bootstrap difference of means, sedan-hatchback**



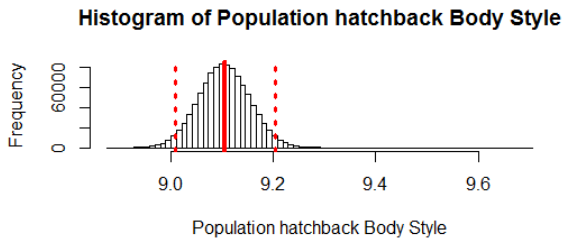
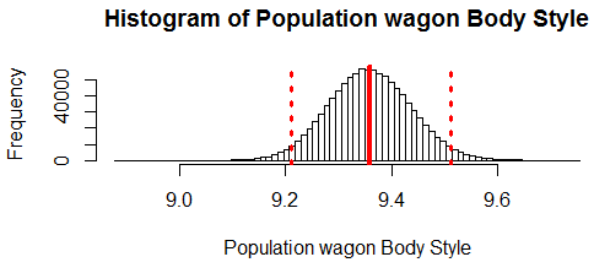


Table 7.9: Bootstrap difference of means, wagon-hatchback



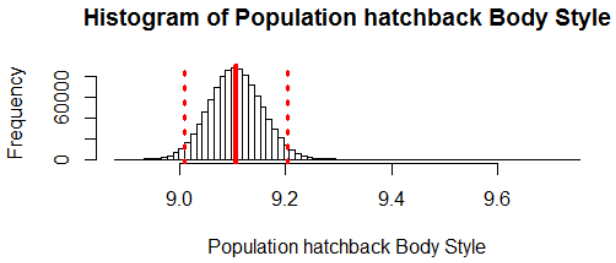


Table 7.10: Bootstrap difference of means, wagon-sedan