Business Analytics using Statistical Modeling
Assignment 5

Question 1

Let's compare the mean load times of Alentus versus HostMonster using their 2 samples

```
library(data.table)
```

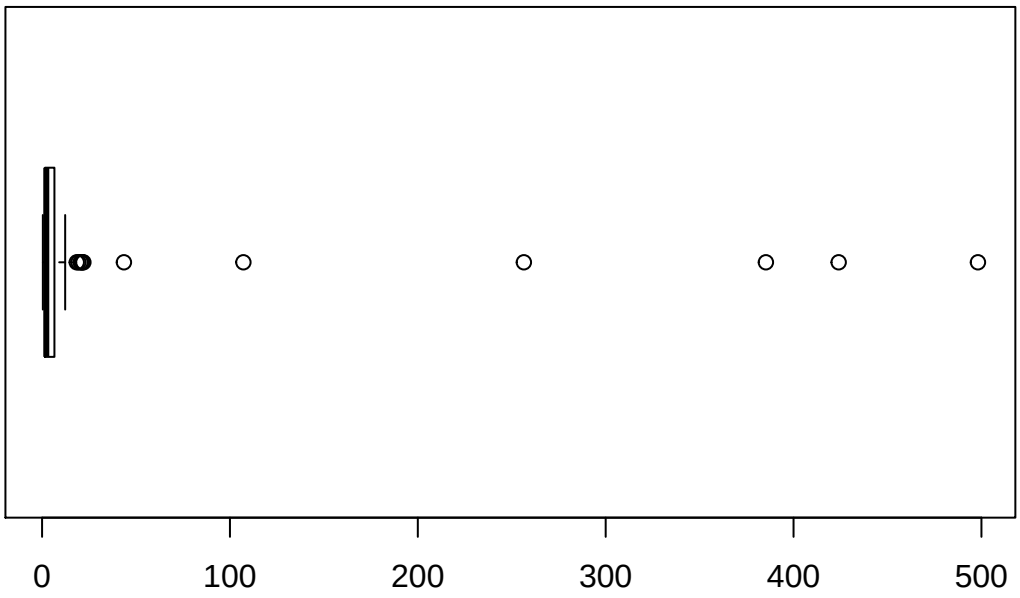
```
## Warning: package 'data.table' was built under R version 3.2.5
```

```
page_loads <- fread('../5-page_loads.csv')
```

CLAIM: Imagine Alentus claims that their mean load time is actually quite comparable to that of HostMonster, if we remove the major outliers from the Alentus load times

a. Use a boxplot to remove the major outliers of Alentus' load times.

```
alentus_box <- boxplot(page_loads$Alentus, horizontal = TRUE)
```




```
alentus_box$out
```

```
##  [1] 498.15 424.01  19.36  21.99 256.45  21.45 385.27  20.49  43.57  18.46
## [11]  21.14 107.15
```

```
alentus_omit_outliers <- page_loads$Alentus[!(page_loads$Alentus %in% alentus_box$out)]
hostmonster <- as.numeric(na.omit(page_loads$HostMonster))
```

Then, use the appropriate form of the `t.test` function to test the difference between the mean of Alentus and the mean of HostMonster load times (assume the sample come from populations with different variances).

```
t_test <- t.test(alentus_omit_outliers, hostmonster, var.equal = FALSE)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  alentus_omit_outliers and hostmonster
## t = 1.9876, df = 132.39, p-value = 0.04892
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.003159185 1.306392097
## sample estimates:
## mean of x mean of y
##  3.177692  2.522917
```

From the output of `t.test`:

i. What is the null and alternative hypotheses in this case?

H_0: $\mu_{\text{alentus\_omit\_outliers}} = \mu_{\text{hostmonster}}$ vs. H_a: $\mu_{\text{alentus\_omit\_outliers}} \neq \mu_{\text{hostmonster}}$

ii. What is the 95% CI of the difference of the 2 providers' means?

```
t_test$conf.int[c(1, 2)]
```

```
## [1] 0.003159185 1.306392097
```

iii. Based on the 95% CI, the t-value, and the p-value, would you reject the null hypothesis or not?

```
t_value <- t_test$statistic
t_value
```

```
##        t
## 1.987635
```

```
p_value <- t_test$p.value
p_value
```

```
## [1] 0.04891539
```

Since the 95% CI does not contain 0, the t-value is larger than 1.96, and the p-value is smaller than 0.05, we can reject the null hypothesis.

b. Let's try using bootstrapping:

Estimate bootstrapped alternative values of t using the same `t.test` function as above to compare bootstrapped samples of both providers; estimate bootstrapped null values of t by using the `t.test` function above to compare bootstrapped values of Alentus against the original Alentus sample; also estimate the difference between means of both bootstrapped samples.

```
bootstrap_null_alt <- function(sample0, sample1) {
  resample0 <- sample(sample0, length(sample0), replace = TRUE)
  resample1 <- sample(sample1, length(sample1), replace = TRUE)
  resample0_se <- sd(resample0) / sqrt(length(resample0))
  t_stat_alt <- t.test(resample0, resample1, var.equal = FALSE)$statistic
  t_stat_null <- t.test(resample0, sample0, var.equal = FALSE)$statistic
  mean_diffs <- abs(mean(resample0) - mean(resample1))
  return(c(t_stat_alt, t_stat_null, mean_diffs))
}
boot_t_stats <- replicate(10000, bootstrap_null_alt(alentus_omit_outliers,
                                                    hostmonster))
```

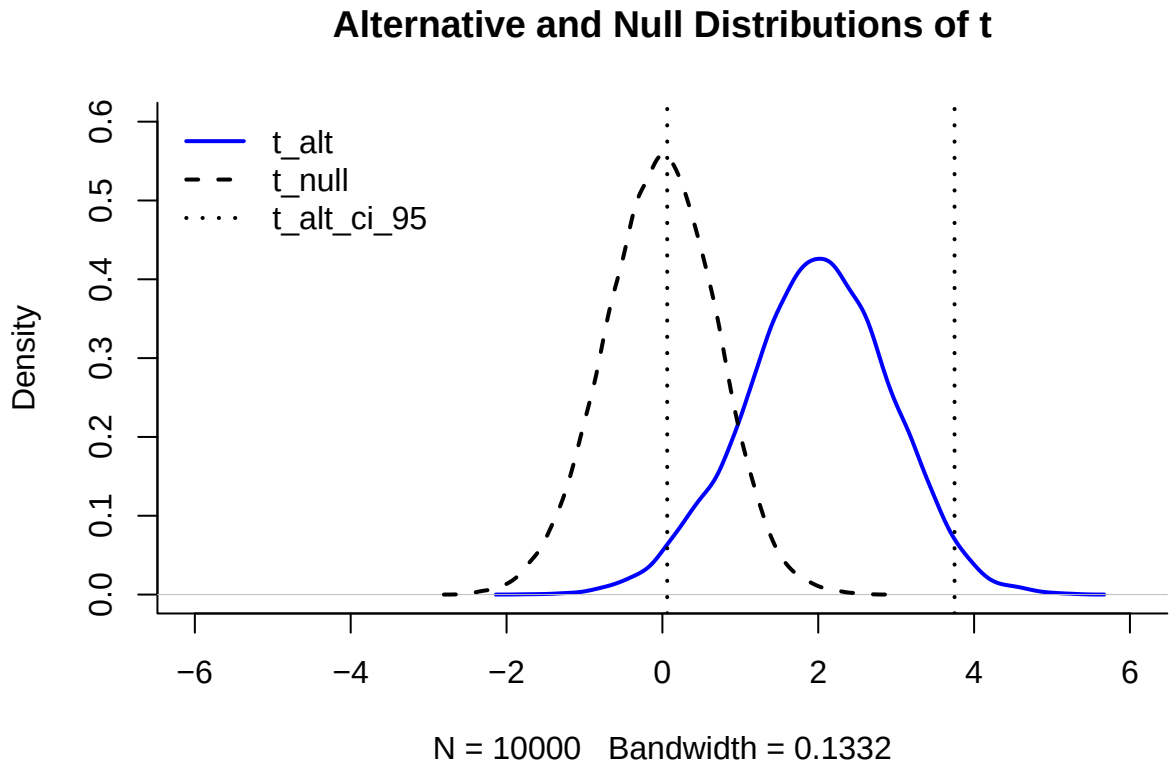
i. What is the bootstrapped 95% CI of the difference of means?

```
mean_diffs <- boot_t_stats[3, ]
mean_diffs_ci_95 <- quantile(mean_diffs, probs = c(0.025, 0.975))
mean_diffs_ci_95
```

```
##       2.5%      97.5%
## 0.06909215 1.32021855
```

ii. Plot a distribution of the bootstrapped null t-values and bootstrapped alternative t-values, adding vertical lines for the 95% CI of the alternative distribution

```
t_alt <- boot_t_stats[1, ]
t_null <- boot_t_stats[2, ]
t_alt_ci_95 <- quantile(t_alt, probs = c(0.025, 0.975))
plot(density(t_alt), xlim = c(-6, 6), ylim = c(0, 0.6), lwd = 2, col = 'blue', bty = 'l',
     main = 'Alternative and Null Distributions of t')
lines(density(t_null), lty = 2, lwd = 2)
abline(v = t_alt_ci_95, lty = 3, lwd = 2)
legend('topleft', legend = c('t_alt', 't_null', 't_alt_ci_95'), lwd = c(2, 2, 2),
       lty = c(1, 2, 3), bty = 'n', col = c('blue', 'black', 'black'))
```



iii. Based on these bootstrapped results, should we reject the null hypothesis?

The `mean_diffs_ci_95` still does not contain 0, so we can reject the null hypothesis.

Question 2

CLAIM: Alentus claims that, with its major outliers removed, its median load time is in fact significantly smaller than the median load time of HostMonster (with 95% confidence)

a. First, confirm that the median load time of Alentus (without outliers) is smaller than for HostMonster

```
median(alentus_omit_outliers) < median(hostmonster)
```

```
## [1] TRUE
```

b. Bootstrap the difference between the median of Alentus (without major outliers) and the median for HostMonster; also bootstrap the 'null' difference (compare the median of bootstrapped samples of Alentus against the median of the original Alentus sample)

```
boot_median_diffs <- function(sample0, sample1) {
  resample0 <- sample(sample0, length(sample0), replace = TRUE)
  resample1 <- sample(sample1, length(sample1), replace = TRUE)
  median_diff <- median(resample0) - median(resample1)
  median_null_diff <- median(resample0) - median(sample0)
  return(c(median_diff, median_null_diff))
}
median_diffs <- replicate(10000, boot_median_diffs(alentus_omit_outliers, hostmonster))
```

i. What is the average difference between medians of the 2 service providers?

```
median_diff <- median_diffs[1, ]
mean(median_diff)
```

```
## [1] -0.220118
```

ii. What is the 95% CI of the difference between the medians of the 2 service providers?

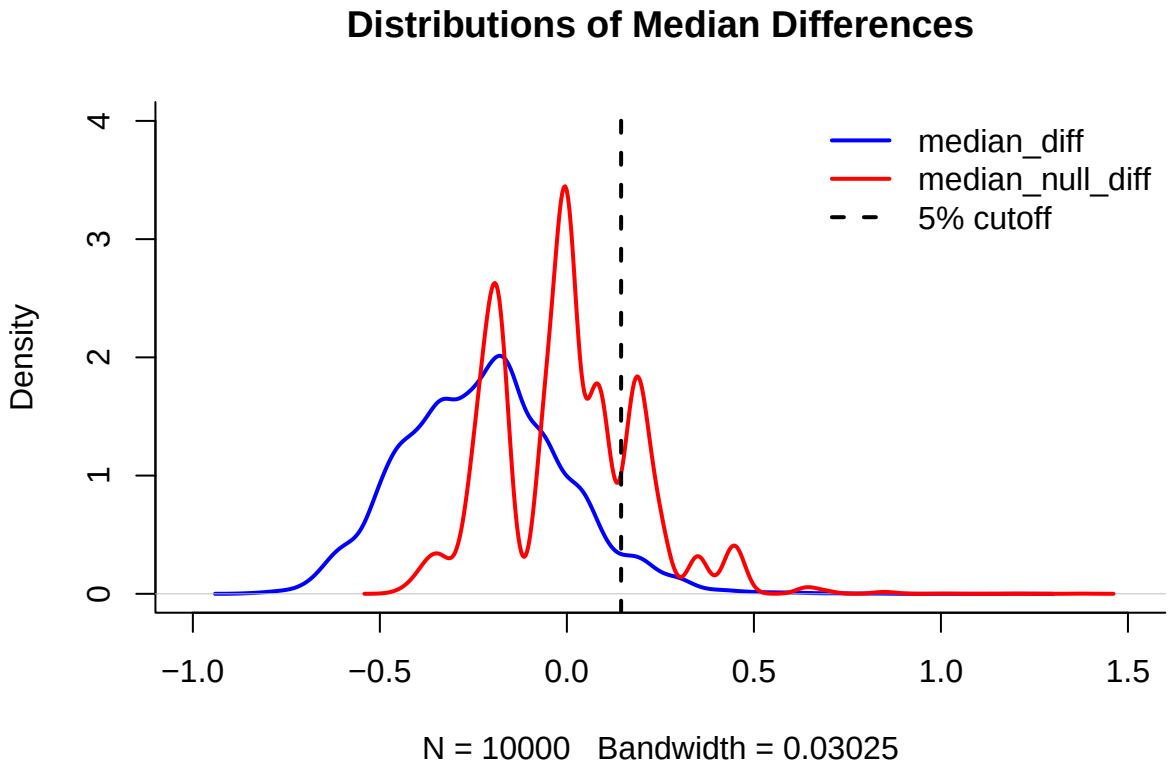
Since we have a `less than` alternative, we only care about the left tail

```
median_diff_ci_95 <- quantile(median_diff, probs = c(0.95))
median_diff_ci_95
```

```
##   95%
## 0.145
```

iii. Plot the distributions of the bootstrapped alternative and null differences between medians and use a vertical dashed lines to show us the 5% rejection zone

```
median_null_diff <- median_diffs[2, ]
plot(density(median_diff), xlim = c(-1, 1.5), ylim = c(0, 4), lwd = 2, col = 'blue',
     bty = 'l', main = 'Distributions of Median Differences')
lines(density(median_null_diff), lwd = 2, col = 'red')
abline(v = median_diff_ci_95, lwd = 2, lty = 2)
legend('topright', legend = c('median_diff', 'median_null_diff', '5% cutoff'),
       lwd = c(2, 2, 2), lty = c(1, 1, 2), col = c('blue', 'red', 'black'), bty = 'n')
```



c. Does the 95% CI bootstrapped difference of medians suggest that the median of Alentus load times (without outliers) is significantly smaller than the median load times of HostMonster?

No. The 95% CI bootstrapped difference of medians still contains values more than 0, so it does not suggest that the median of Alentus load times (without outliers) is significantly smaller than the median load times of HostMonster.

Question 3

Let's take a look back at some data from a marketing survey of mobile users. In particular, we are interested in responses by iPhone versus Samsung users to a brand identification question.

```
survey <- fread('../5-Data_0630.txt')
iphone <- survey[survey$`[current_phone]` == 1, ]$`[Brand_Identification.1]`
samsung <- survey[survey$`[current_phone]` == 2, ]$`[Brand_Identification.1]`
```

We find that the means of identification scores between users of the 2 phone brands are very similar. So we wish to test whether 1 brand's variance of identification scores is higher than the other brand's variance of identification scores.

Start by identifying which brand has the higher variance.

```
which.max(c(var(iphone), var(samsung)))
```

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

The iPhone has the higher variance

a. What is the null and alternative hypotheses in this case?

H_0: $\sigma_{\text{iphone}} \leq \sigma_{\text{samsung}}$ vs. H_a: $\sigma_{\text{iphone}} > \sigma_{\text{samsung}}$

b. Let's try traditional statistical methods first:

i. What is the F-statistic of the ratio of variances?

```
f_value <- var(iphone) / var(samsung)
f_value
```

```
## [1] 1.136295
```

ii. What is the cut-off value of F, such that we want to reject the 5% most extreme F-values?

```
cut_off <- qf(p = 0.95, df1 = length(iphone) - 1, df2 = length(samsung) - 1)
cut_off
```

```
## [1] 1.369645
```

iii. Can we reject the null hypothesis?

Since `f_value` is less than the `cut_off`, we do not reject the null hypothesis.

c. Let's try bootstrapping this time

i. Create bootstrapped values of the F-statistic for both null and alternative hypotheses.

```
var_brands_test <- function(larger_var_sample, smaller_var_sample) {
  resample_larger_var <- sample(larger_var_sample, length(larger_var_sample),
                                replace = TRUE)
  resample_smaller_var <- sample(smaller_var_sample, length(smaller_var_sample),
                                 replace = TRUE)
  f_alt <- var(resample_larger_var) / var(resample_smaller_var)
  f_null <- var(resample_larger_var) / var(larger_var_sample)
  return(c(f_alt, f_null))
}
f_stats <- replicate(10000, var_brands_test(iphone, samsung))
f_alts <- f_stats[1, ]
f_nulls <- f_stats[2, ]
```

ii. What is the 95% cutoff value according to the bootstrapped null values of F?

```
boot_cut_off <- quantile(f_nulls, probs = 0.95)
boot_cut_off
```

```
##      95%
## 1.183602
```

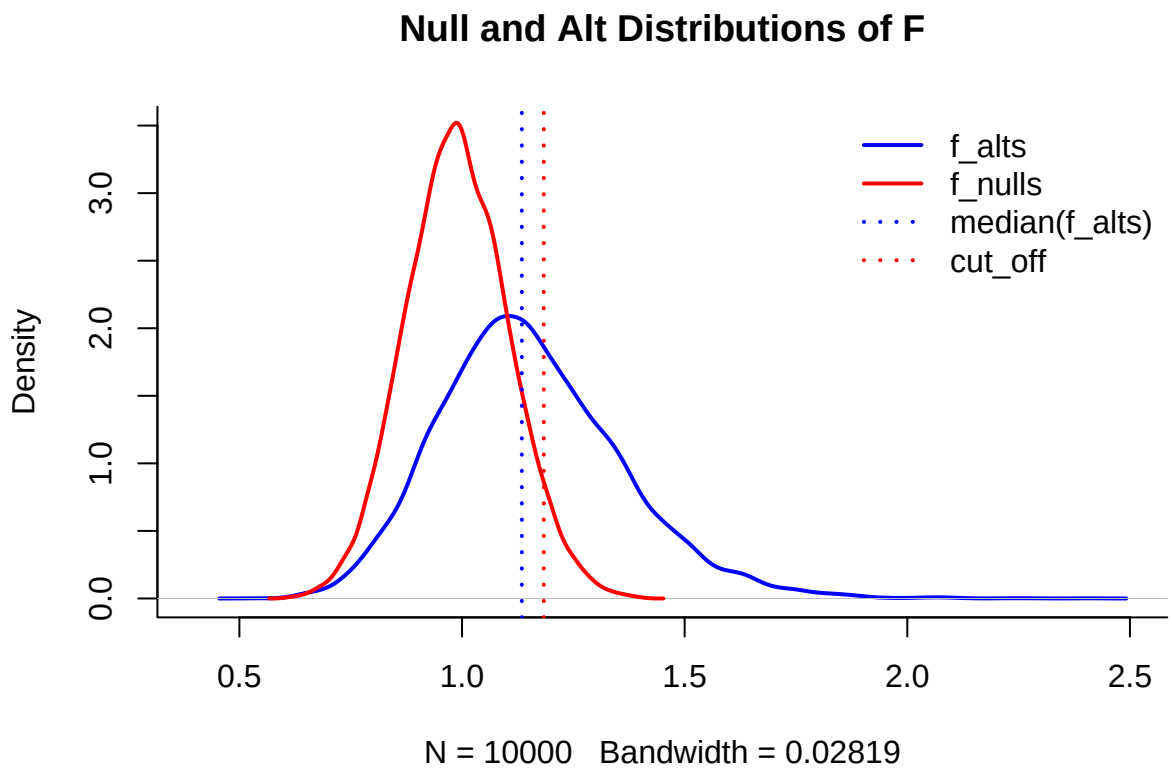
iii. What is the median bootstrapped F-value for the alternative hypothesis?

```
median(f_alts)
```

```
## [1] 1.134318
```

iv. Plot a visualization of the null and alternative distributions of the bootstrapped F-statistic, with vertical lines at the cutoff value of F nulls, and at median F-values for the alternative.

```
plot(density(f_alts), xlim = c(0.4, 2.5), ylim = c(0, 3.5), lwd = 2, col = 'blue',
     bty = 'l', main = 'Null and Alt Distributions of F')
lines(density(f_nulls), lwd = 2, col = 'red')
abline(v = boot_cut_off, lty = 3, lwd = 2, col = 'red')
abline(v = median(f_alts), lty = 3, lwd = 2, col = 'blue')
legend('topright', legend = c('f_alts', 'f_nulls', 'median(f_alts)', 'cut_off'),
       lty = c(1, 1, 3, 3), lwd = c(2, 2, 2, 2), bty = 'n',
       col = c('blue', 'red', 'blue', 'red'))
```



v. What do the bootstrap results suggest about the null hypothesis?

The median F-value for the alternative is less than the cutoff value of F nulls, so we do not reject the null hypothesis.