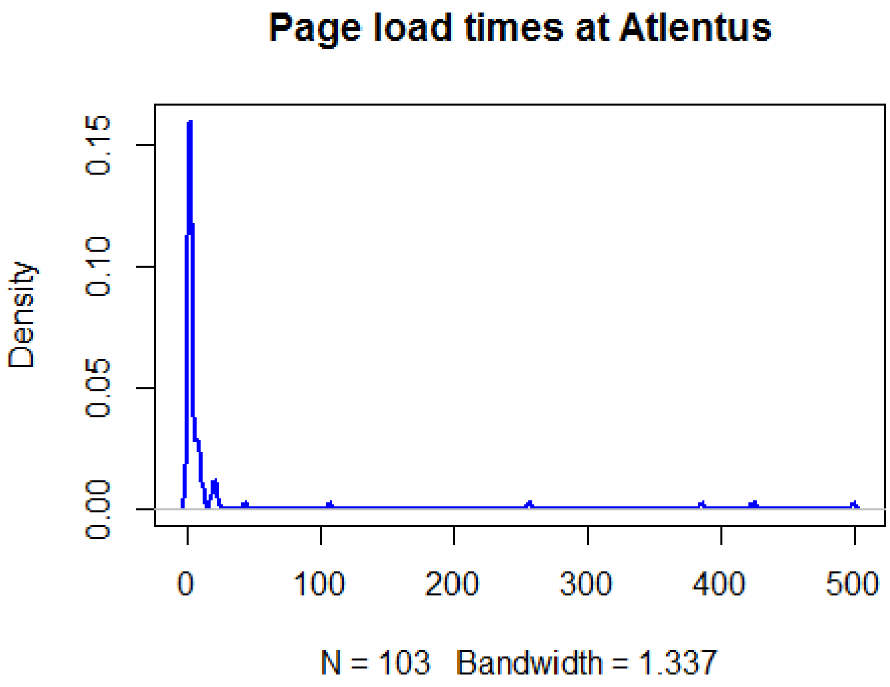
HW5 (answer are in blue text)

Question 1) Let's compare the mean load times of Alentus versus HostMonster using their two samples (see class notes on difference of means between two samples; data in page_loads.csv)

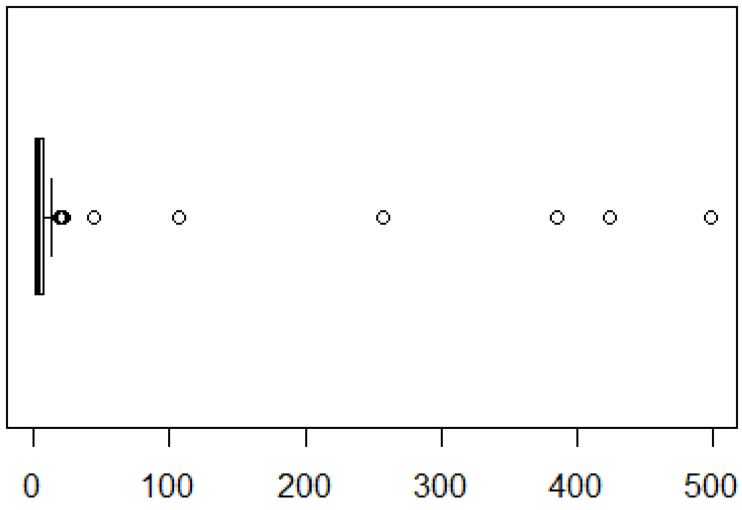
```
page_loads <- read.csv("c:/Users/tsunh/Desktop/schoolwork/BASM/page_loa
ds.csv")
alentus <- page_loads$Alentus
hostmonster <- page_loads$HostMonster
```

a)Use a boxplot to remove the major outliers of Alentus' load times. 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 samples come from populations with different variances). From the output of t.test:

```
plot(density(alentus), lwd=2, col='blue', main="Page load times at Atle
ntus")
```



```
boxplot(alentus, horizontal = TRUE)
```



```
su <- summary(alentus)
su

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.41    1.55    2.24   20.65    6.58  498.20

Q1 <- su["1st Qu."]
Q3 <- su["3rd Qu."]
IQR <- Q3 - Q1

#get rid of outlier; use filter function from dplyr package
library(dplyr)

alentus_cleaned <- page_loads %>%
  filter(Alentus > Q1-1.5*IQR & Alentus < Q3+1.5*IQR) %>%
  select(Alentus)
alentus_cleaned <- alentus_cleaned$Alentus

#omit na in hostmonster
hostmonster_cleaned <- na.omit(hostmonster)
mean_hostmonster <- mean(hostmonster_cleaned)

#t test
alentus_t_cutoff <- abs(qt(0.025, df=length(alentus_cleaned)-1))
alentus_t_cutoff

## [1] 1.986675
```

```r
mean_diff <- mean(alentus_cleaned)- mean_hostmonster
t.test(hostmonster_cleaned, alentus_cleaned, alternative="two.sided")

##
##  Welch Two Sample t-test
##
## data:  hostmonster_cleaned and alentus_cleaned
## 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:
##  -1.306392097 -0.003159185
## sample estimates:
## mean of x mean of y
##  2.522917  3.177692
```

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

Here we are going to test the **difference** between the mean of Alentus and the mean of HostMonster load times,2.522917. Therefor, our hypotheses are shown below:

Let $\mu_0 = mean\ of\ alentus$, and $\mu_1 = mean\ of\ hostmonster$

$H_{null}: \mu_0 - \mu_1 = 0$

$H_{alt}: \mu_0 - \mu_1 \neq 0$

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

The 95% CI of the difference of the two providers' means is (-1.306392097,

-0.003159185), which 0 is not included.

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

1. The 95% CI didn't include zero.
2. The t-value is greater than the t cutoff. |-1.9876| > 1.986675
3. The p-value is smaller than significance level. 0.04892 < 0.05

Therefore, I would like to reject the null hypothesis and claim that those means are different!

b)Let's try this 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.

```r
set.seed(38)
bootstrap_null_alt <- function(sample0, hyp_mean) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  resample_se <- sd(resample) / sqrt(length(resample))
  mean_diff <- mean(resample)-hyp_mean
  t_stat_alt <- (mean(resample) - hyp_mean) / resample_se
  t_stat_null <- (mean(resample) - mean(sample0)) / resample_se
  return(c(t_stat_alt, t_stat_null,mean_diff))
}
boot_t_stats <- replicate(2000, bootstrap_null_alt(alentus_cleaned, mea
n_hostmonster))
t_alt <- boot_t_stats[1,]
t_null <- boot_t_stats[2,]
mean_diff <- boot_t_stats[3,]
```

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

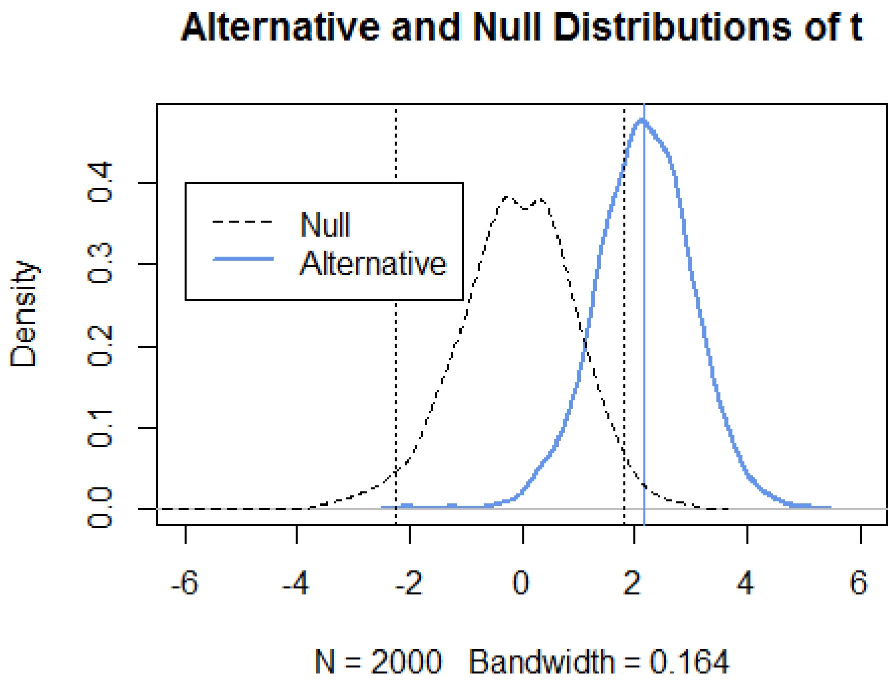
```
ci_95 <- quantile(mean_diff, probs=c(0.025, 0.975))
ci_95

##      2.5%     97.5%
## 0.1155284 1.2503086
```

The bootstrapped 95% CI of the difference of means is (0.1155284, 1.2503086). Note that the interval doesn't include zero!

ii. Plot a distribution of the bootstrapped null t-values and bootstrapped alternative t-values, adding vertical lines for the 95% CI of the null distribution (adjust x- and y- axis limits accordingly).

```
plot(density(t_alt), xlim=c(-6,6), lwd=2,main="Alternative and Null Dis
tributions of t", col="cornflowerblue")
abline(v=mean(t_alt), col="cornflowerblue")
lines(density(t_null), lty="dashed")
ci_95 <- quantile(t_null, probs=c(0.025, 0.975))
abline(v=ci_95, lty="dotted")
legend(-6,0.4,c("Null","Alternative"),lty = c(2,1),lwd=c(1,2),col = c("
black","cornflowerblue"))
```



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

Our alternative t-values lie outside the 95% CI of the null t-distribution. Therefore, we should reject the null hypothesis. The two mean should be difference

Question 2) Here, we really don't know what test statistic to use to compare medians of two samples, so let's just bootstrap the confidence interval:

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)!

$H_{null}: Median_{Alentus} = Median_{Hostermonster}$

$H_{alt}: Median_{Alentus} < Median_{Hostermonster}$

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

```
median(alentus_cleaned) < median(hostmonster_cleaned)

## [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).

```
set.seed(38)
# function for bootstrapping f distribution
bootstrap_median_null_alt <- function(sample0, hyp_median) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  median_diff <- median(resample)-hyp_median
  t_stat_alt <- (median(resample) - hyp_median)
  t_stat_null <- (median(resample) - median(sample0))
  return(c(t_stat_alt, t_stat_null,median_diff))
}
boot_t_stats <- replicate(2000, bootstrap_median_null_alt(alentus_clean
ed, median(hostmonster_cleaned)))
median_alt <- boot_t_stats[1,]
median_null <- boot_t_stats[2,]
median_diff <- boot_t_stats[3,]
```

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

```
mean(median_diff)

## [1] -0.21008
```

ii. What is the 95% CI of the difference between the medians of the two service providers? (consider this time, where the 5% ‘rejection zone’ of the distribution of differences will be: will it be on both sides or only on one side?)

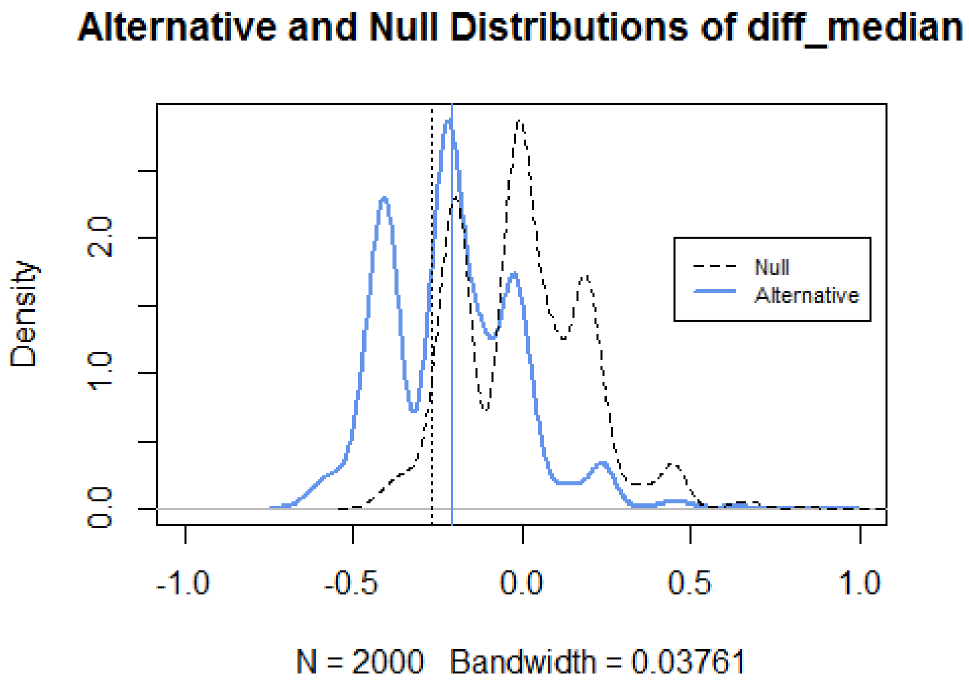
```
ci_95 <- quantile(median_diff,probs = c(0.5))
ci_95

##   50%
## -0.21
```

According to the alternative hypothesis, we could use left tailed test, whose 5% rejection zone will show in one side (left). The 95% CI of the difference of the two medians is (-0.21, ∞)

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’

```
plot(density(median_alt), xlim=c(-1,1), lwd=2,main="Alternative and Nul
l Distributions of diff_median", col="cornflowerblue")
abline(v=mean(median_alt), col="cornflowerblue")
lines(density(median_null), lty="dashed")
ci_95 <- quantile(median_null, probs=c(0.05))
abline(v=ci_95, lty="dotted")
legend(0.45,2,c("Null","Alternative"),lty = c(2,1),lwd=c(1,2),col = c("
black","cornflowerblue")
       ,cex = 0.7)
```



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?

From the above plot, we can find out that the average of median-difference lies in the 95% CI instead of rejection zone. Therefore, we are not going to reject the null hypothesis; that is, the median of Alentus load time is not significantly smaller than that of HostMonster.

Question 3) Let's take a look back at some data from a marketing survey of mobile users.

```
#load data
survey <- read.csv("c:/Users/tsunh/Desktop/schoolwork/BASM/Data_0630.tx
t", sep="\t", header = TRUE)
#get column we are interested in
iphone <- survey[survey$X.current_phone.==1,]$X.Brand_Identification.1.
samsung <- survey[survey$X.current_phone.==2,]$X.Brand_Identification.1.
```

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

a. What is the null and alternative hypotheses in this case?(Start by identifying which brand has the higher variance)

```
var(iphone)-var(samsung)

## [1] 0.3480833
```

After checking whose variance might be bigger, we determined to test whether iphone's variance of identification scores is higher than that of samsung's. The the null and alternative hypotheses in this case would be:

$H_{null}: \mathrm{Var}_{iphone} = \mathrm{Var}_{samsung}$

$H_{alt}: \mathrm{Var}_{iphone} > \mathrm{Var}_{samsung}$

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? (this is another way of saying we want 95% confidence) Use the qf() function in R to determine the cutoff.

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

## [1] 1.369645
```

iii. Can we reject the null hypothesis?

According to the tranditional statistics, we should not reject the null hypothesis. Also, we can use *var.test* function.

```
var.test(iphone,samsung,alternative = "greater")

## 
##  F test to compare two variances
## 
## data:  iphone and samsung
## F = 1.1363, num df = 108, denom df = 112, p-value = 0.2516
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.8296279       Inf
## sample estimates:
## ratio of variances 
##           1.136295
```

Note that the p-value is 0.2516 which is greater than 0.05, so we are not going to reject the null hpothesis. The two variance should be the same accordingly.

c. Let’s try bootstrapping this time:

```r
# function for bootstrapping f distribution
set.seed(38)
var_phone_test <- function(larger_var_sample,smaller_var_sample){
  resample_larger_var <- sample(larger_var_sample,length(larger_var_sam
ple),replace = T)
  resample_smaller_var <- sample(smaller_var_sample,length(smaller_var_
sample),replace = T)
  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))
}
```

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

```r
f_stats <- replicate(5000,var_phone_test(iphone,samsung))
f_alts <- f_stats[1,]
f_nulls <- f_stats[2,]
```

The bootstrapped values of the F-statistic are stored in `f_alts` and `f_nulls`.

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

```r
quantile(f_nulls,probs = 0.95)

##      95%
## 1.185212
```

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

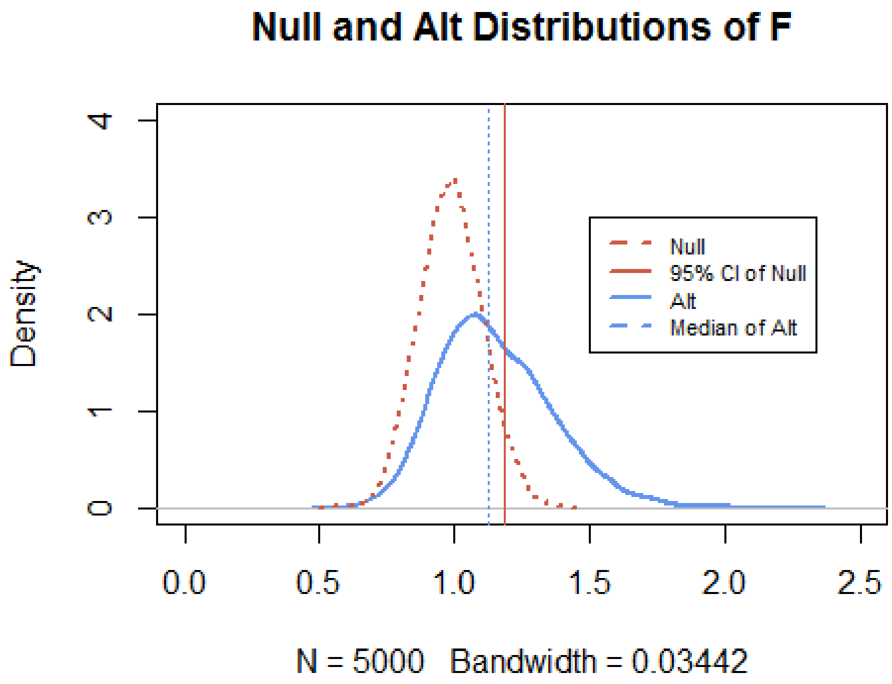
```r
median(f_alts)

## [1] 1.127677
```

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.

```r
plot(density(f_alts),col="cornflowerblue",ylim = c(0,4),xlim=c(0,2.5),m
ain="Null and Alt Distributions of F", lwd = 2)
lines(density(f_nulls),col="coral3",lwd=2,lty="dotted")

abline(v=quantile(f_nulls,probs=0.95),col="coral3")
abline(v=median(f_alts),lty="dotted",col="cornflowerblue")
legend(1.5,3,c("Null","95% CI of Null","Alt","Median of Alt"),
       lty=c(2,1,1,2),
       col = c("coral3","coral3","cornflowerblue","cornflowerblue"),
       lwd = c(2,2,2,2),
       cex=0.7)
```



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

The boostrap results suggest that we should not reject the null hypothesis since the median of alternative distribution lies into the 95% CI of the null distribution. Therefore, there's no significant difference between the variances.