Bootstrapping Example

Yu Han Huang

> Problem One: The Hsinchu Rubber Company is in financial trouble because of a reputation of poor tire quality. They have launched a new advertising campaign to change their image. In their new advertisement, they claim that their tires can run an average of 90,000 km before needing to be replaced. The editors of a Taiwanese consumer magazine are skeptical and wanted to test the claim. They collected data from 360 drivers who use these tires and recorded how long the tires lasted. On average, the tires in this sample lasted 85945.29 km with a standard deviation of 14996.55 km. Is the company's new advertising claim to
> be believed? Use bootstrap methods to examine this problem.

We first read in the data to our code. (load tires.csv file) and set the numbers of bootstrapping to 2000

```
tire <- read.csv(file.choose(),header = TRUE)
num_bootstraps <- 2000
hypo_mean <- 90000
```

Calculate 95% Confidence Interval

```
#Function
ninetyfivec <- function(smean,ssde){
  paste("(",smean-1.96*ssde,",",smean+1.96*ssde,")")
}
```

Estimation of population mean We can use our sample mean to estimate the population mean.

```
mean(tire$lifetime_km)
```

```
## [1] 85945.29
```

Since the sample mean is the estimated approximation of population mean, we get that the estimated population mean is 85945.29.

Bootstrapped 95% Confidence Interval of the estimation of population mean

```
#Function
boot_mean <- function(sample0) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  return(mean(resample))
}

#Calculation
bootmean <- replicate(num_bootstraps,boot_mean(tire$lifetime_km))
ninetyfivec(mean(bootmean),sd(bootmean))
```

```
## [1] "( 84450.7556699178 , 87451.0172777488 )"
```

Mean Bootstrapped Difference The difference should be zero if null hypothesis is correct.

```
#Function
boot_mean_diffs <- function(sample0, mean_hyp) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  return( mean(resample) - mean_hyp )
}

#Calculation
```

```
mean_diffs <- replicate(num_bootstraps,boot_mean_diffs(tire$lifetime_km,hypo_mean))
mean(mean_diffs)
```

```
## [1] -4029.252
```

Mean Bootstrapped Difference - 95% Confidence Interval

```
quantile(mean_diffs, probs=c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -5656.581 -2417.079
```

Mean Bootstrapped T-Statistics

```
#Function
boot_t_stat <- function(sample0, mean_hyp) {
  resample <- sample(sample0, length(sample0), replace=TRUE)
  diff <- mean(resample) - mean_hyp
  resample_se <- sd(resample)/sqrt(length(resample))
  return( diff/resample_se )
}

#Calculation
t_boots <- replicate(num_bootstraps, boot_t_stat(tire$lifetime_km, hypo_mean))
mean(t_boots)
```

```
## [1] -5.172596
```

Mean Bootstrapped T-Statistics - 95% Confidence Interval

```
quantile(t_boots, probs=c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -7.241169 -3.163598
```

From our calculation, we see that under the 5% significance level, we need to reject the null hypothesis since our bootstrapped t-score exceeds our critical value. Thus, we reject the null hypothesis and show that the advertising claim from the company is not true.

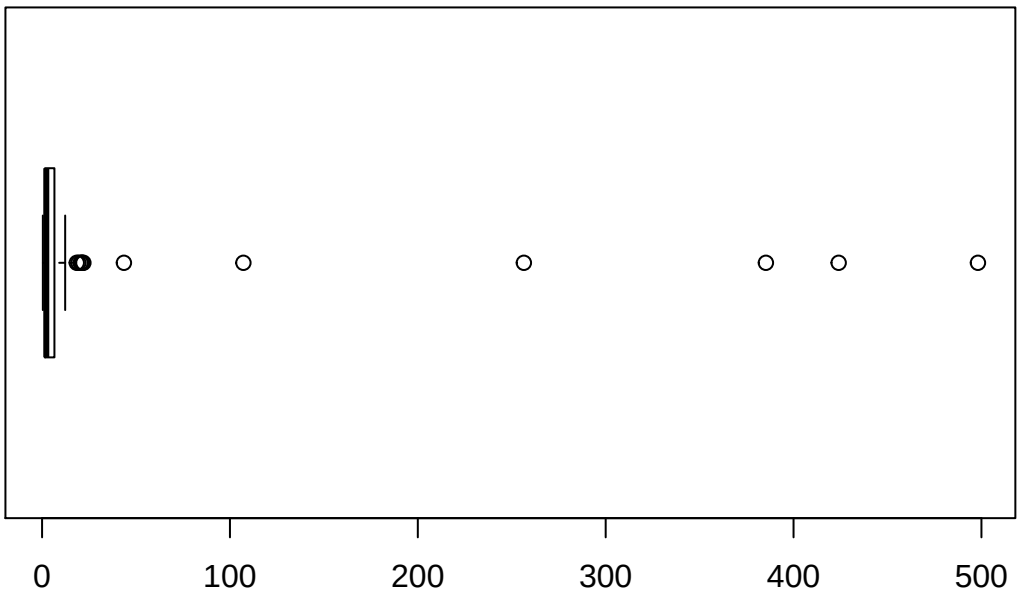
> Problem Two: In February of 2011 I received an email from an irritated customer which indicated that my website (www.SigmaZone.com) was very slow. The most obvious solution to a slow site would be to notify my web hosting company, Alentus. I did some preliminary testing and it did appear that my site was not only slower than big sites such as Google and Corel, but was actually much slower than Alentus' home page. 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. Test on Alentus's claim using bootstrap.

Load file page_loads.csv and set bootstrap times to 2000

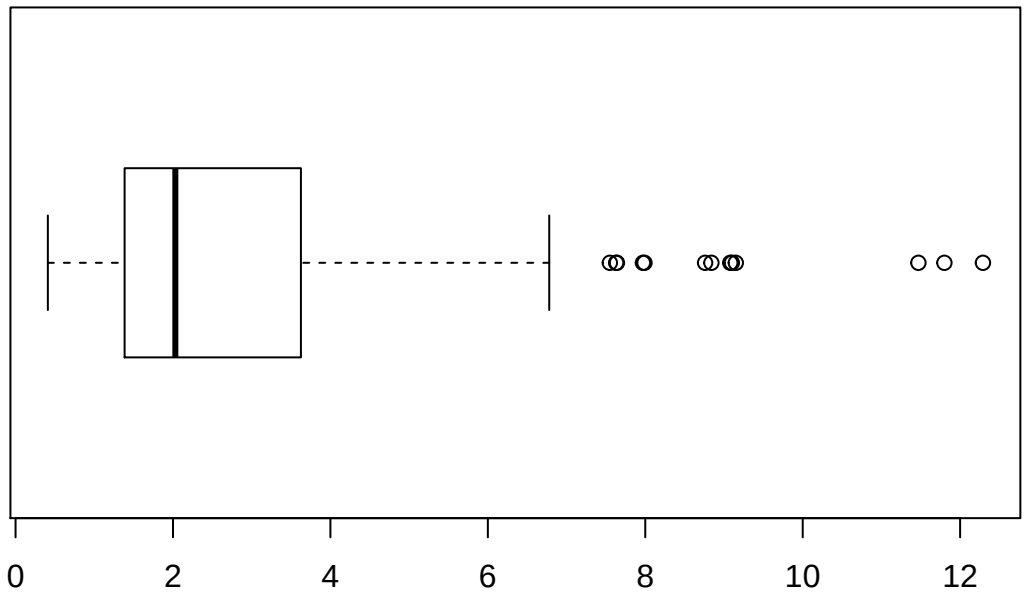
```
load_times <- read.csv(file.choose(), header = TRUE)
host <- load_times$HostMonster
host_clear <- na.omit(host)
num_bootstraps <- 2000
```

Use a boxplot to remove the major outliers

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




```
alentus_omit_outliers <- alentus[!(alentus %in% alentus_box$out)]
boxplot(alentus_omit_outliers, horizontal = TRUE)$out
```




```
##  [1] 11.80  8.84  7.99  7.97  9.15  9.08  9.10  8.76 12.29  7.64  7.55
## [12]  7.63 11.47
```

Estimate bootstrapped alternative values of t to compare bootstrapped samples of both providers Alternative t-value: Standardized difference between bootstrapped and hypothesized mean

```
#Function
bootstrap_alt_t <- function(sample0,sample1){
  resample_0 <- sample(sample0,size = length(sample0),replace = TRUE)
  resample_1 <- sample(sample1,size = length(sample1),replace = TRUE)
  alt_t <- t.test(resample_0,resample_1,var.equal = FALSE)
  return(alt_t$statistic)}

#Calculation
boot_alt_t <- replicate(num_bootstraps,bootstrap_alt_t(alentus_omit_outliers, host_clear))
```

Estimate bootstrapped null values of t to compare bootstrapped values of Alentus against the original Alentus sample Null t-value: What t-values would be if original sample mean was hypothesized; Difference between bootstrap mean and original sample mean (should be zero on average, across bootstrapped samples)

```
#Function
bootstrap_null_t <- function(sample0,com){
  resample_0 <- sample(sample0, size = length(sample0), replace = TRUE)
  null_t <- t.test(resample_0, com, var.equal = FALSE)
  return(null_t$statistic)}

#Calculation
boot_null_t <- replicate(num_bootstraps,bootstrap_null_t(alentus_omit_outliers,alentus_omit_outliers))
```

Estimate the difference between means of both bootstrapped samples.

```
#Function
bootstrap_diff_t <- function(sample0,sample1){
  resample_0 <- sample(sample0,size = length(sample0),replace = TRUE)
  resample_1 <- sample(sample1,size = length(sample1),replace = TRUE)
  diff_t <- mean(resample_0)-mean(resample_1)
  return(diff_t)
}

#Calculation
boot_diff_t <- replicate(num_bootstraps,bootstrap_diff_t(alentus_omit_outliers, host_clear))
```

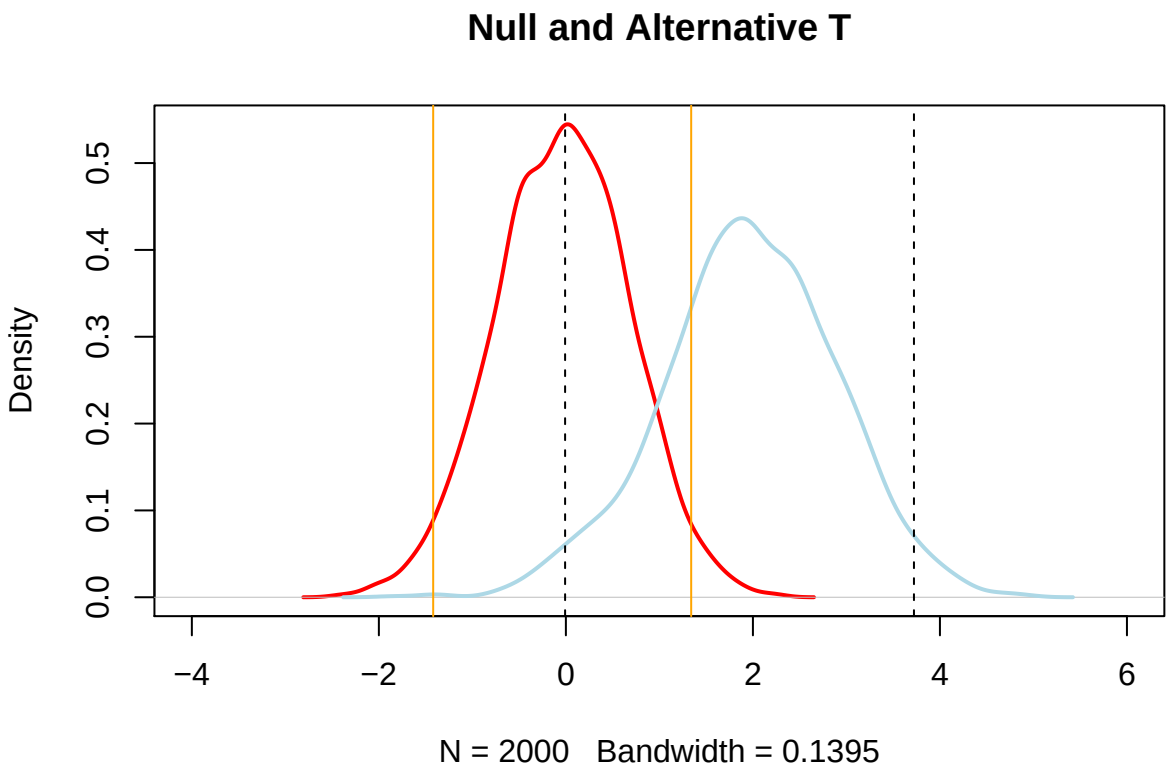
(i) What is the bootstrapped 95% CI of the difference of means?

```
quantile(boot_diff_t, probs=c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.03344989 1.29869059
```

(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 (adjust x- and y- axis limits accordingly)

```
c95_null <- quantile(boot_null_t, probs=c(0.025, 0.975))
c95_alt <- quantile(boot_alt_t, probs=c(0.025, 0.975))
plot(density(boot_null_t), col="red", xlim=c(-4,6), lwd=2, main = "Null and Alternative T")
lines(density(boot_alt_t), col="light blue",xlim=c(-4,6), lwd=2)
abline(v=c95_alt, lty="dashed")
abline(v=c95_null, col="orange")
```



(iii) Based on these bootstrapped results, should we reject the null hypothesis?
From the bootsrapped result, we can find that our alternative hypothesis lie largely outside from the 95% Confidence Interval of Null Hypothesis. And that the 95% Confidence Interval of Differences of mean doesn't contain 0. Hence, we may conclude that we should reject the null hypothesis.

Problem Three: Suppose that Alentus reclaimed that: * 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)!
Test on this claim.

We can bootstrap the difference of medians to find the confidence interval of the difference:

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

```
ifelse(median(alentus_omit_outliers)<median(host_clear), "It is indeed smaller", "False claim! It is no
```

```
## [1] "It is indeed smaller"
```

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).

```
#Function
bootstrap_diff_median <- function(sample0,sample1){
  resample_0 <- sample(sample0,size = length(sample0),replace = TRUE)
  resample_1 <- sample(sample1,size = length(sample1),replace = TRUE)
  diff_median <- median(resample_0)-median(resample_1)
  return(diff_median)
}
```

```r
bootstrap_null_median <- function(sample0,com){
  resample_0 <- sample(sample0, size = length(sample0), replace = TRUE)
  null_median <- median(resample_0)-median(com)
  return(null_median)}

#Calculation
boot_diff_median <- replicate(num_bootstraps,bootstrap_diff_median(alentus_omit_outliers, host_clear))
boot_null_median <- replicate(num_bootstraps,bootstrap_null_median(alentus_omit_outliers, alentus_omit_
```

(b-i) What is the average difference between medians of the two service providers?

```r
mean(boot_diff_median)
```

```
## [1] -0.2219875
```

(b-ii) What is the 95% CI of the difference between the medians of the two service providers?

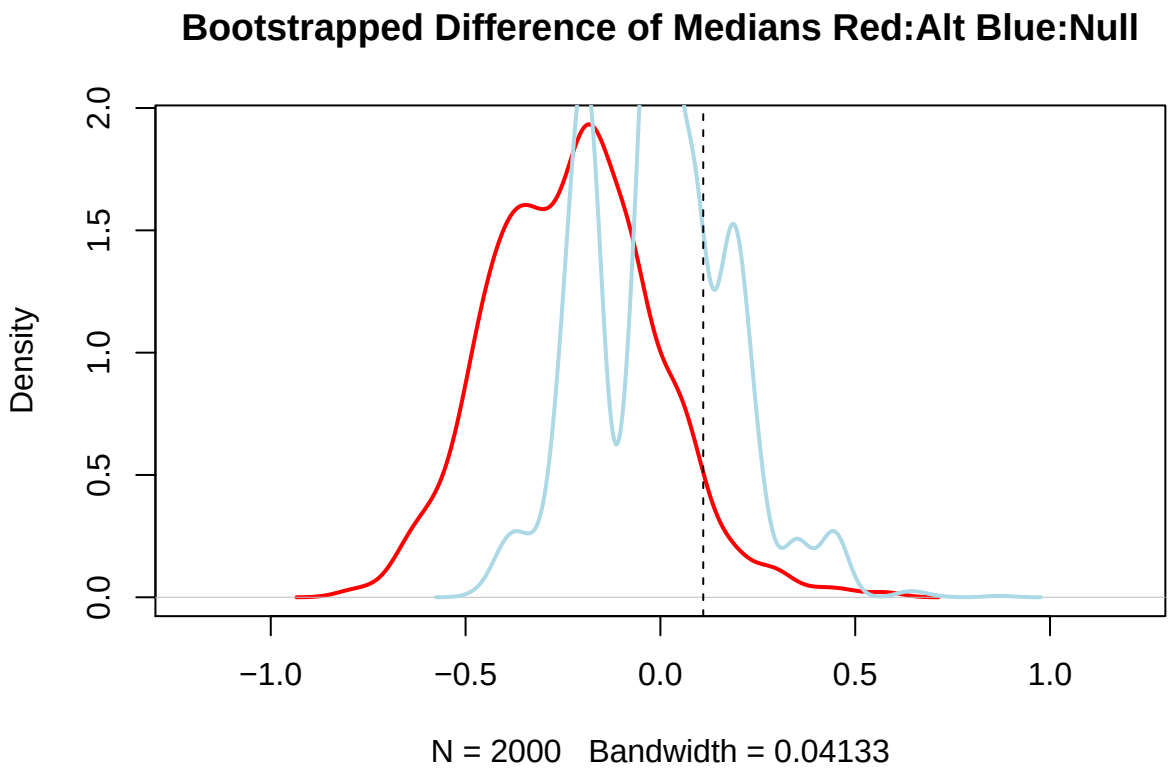
```r
quantile(boot_diff_median,probs = c(0,0.95))
```

```
##    0%   95%
## -0.81  0.11
```

Since here, we are testing whether Alentus' median load time is in fact significantly smaller than the median load time of HostMonster, its H0 should be: boot_diff_median < 0, and that we need to use the right-tail test. Thus our 95% confidence interval should be on one side only, and the rejection zone will be on the right.

(b-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'.

```r
c95_diffmed <- quantile(boot_diff_median, probs=0.95)
plot(density(boot_diff_median), col="red", xlim=c(-1.2,1.2), lwd=2, main = "Bootstrapped Difference of M
lines(density(boot_null_median), col="light blue",xlim=c(-1.2,1.2), lwd=2)
abline(v=c95_diffmed, lty="dashed")
```



Alternative t-distribution: actual difference between sample and hypothesized mean Null t-distribution: expected difference if sample mean equals hypothesized mean

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?
Since 0 is smaller than the critical value and is contained in the 95% CI, we can't reject the null hypothesis that the median of Alentus load times (without outliers) is significantly smaller than the median load times of HostMonster.

Problem Four: We are interested in responses by iPhone versus Samsung users to a brand identification question: "[brand] users share the same values as I do" — [brand] is the respondent's phone brand (iPhone/Samsung) Responses are scored from 1-7 where 1 is "Strongly disagree", 4 is "Neutral", and 7 is "Strongly Agree" 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.

Load file Data_0630.txt

```
survey <- read.csv(file.choose(), sep="\t", header = TRUE)
iphone <- survey[survey$X.current_phone.==1,]$X.Brand_Identification.1.
samsung <- survey[survey$X.current_phone.==2,]$X.Brand_Identification.1.
```

What is the null and alternative hypotheses in this case?

```
var(iphone)
```

```
## [1] 2.901971
```

```
var(samsung)
```

```
## [1] 2.553887
```

Since iphone has a higher variance, the null hypothesis is:

```
"Ho: i-s_variance > 0"
"H1: i-s_variance =< 0"
```

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

```
#Function
bootstrap_alt_f <- function(larger_sd_sample, smaller_sd_sample){
  resample_larger_sd <- sample(larger_sd_sample, length(larger_sd_sample), replace=TRUE)
  resample_smaller_sd <- sample(smaller_sd_sample, length(smaller_sd_sample), replace=TRUE)
  alt_f <- var(resample_larger_sd) / var(resample_smaller_sd)
  return(alt_f)}

bootstrap_null_f <- function(larger_sd_sample){
  resample_larger_sd <- sample(larger_sd_sample, length(larger_sd_sample), replace=TRUE)
  null_f <- var(resample_larger_sd) / var(larger_sd_sample)
  return(null_f)}

#Calculation
boot_alt_f <- replicate(num_bootstraps,bootstrap_alt_f(iphone, samsung))
boot_null_f <- replicate(num_bootstraps,bootstrap_null_f(iphone))
```

(c-i) What is the 95% cutoff value according to the bootstrapped null values of F? *what most F-values should be less than if Null hypothesis is true*

```
cutoff_bootnullf <- quantile(boot_null_f, probs=0.95)
cutoff_bootnullf
```

```
##      95%
## 1.188716
```

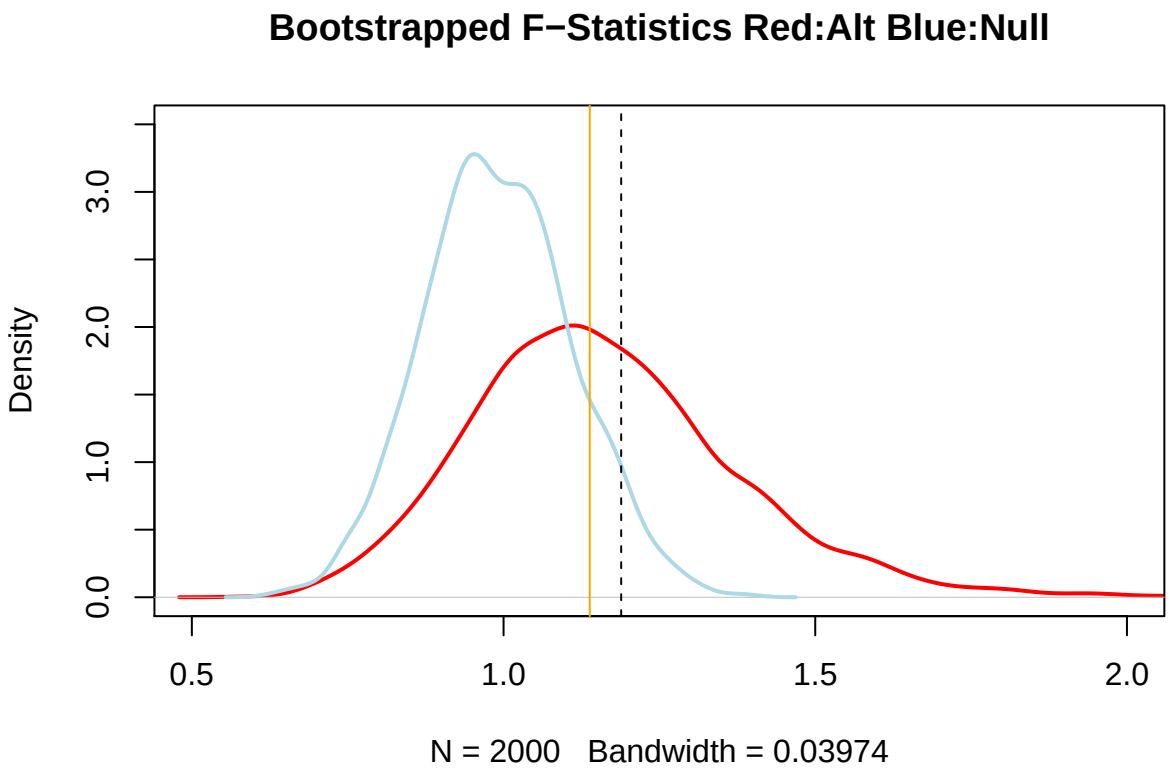
(c-ii) What is the median bootstrapped F-value for the alternative hypothesis? *Our bootstrapped F-value*

```
median_bootaltf <- median(boot_alt_f)
median_bootaltf
```

```
## [1] 1.138264
```

(c-iii) 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(boot_alt_f), col="red", xlim=c(0.5,2), ylim=c(0,3.5),lwd=2, main = "Bootstrapped F-Statisti
lines(density(boot_null_f), col="light blue",xlim=c(0.5,2), ylim=c(0,3.5), lwd=2)
abline(v=cutoff_bootnullf, lty="dashed")
abline(v=median_bootaltf, col="orange")
```



(c-iv) What do the bootstrap results suggest about the null hypothesis?
Since our median of bootstrapped null f is less than the cutoff value, we fail to reject the null hypothesis.