Answer Key to R HW5

Boyie Chen

12/19/2019

A. Bootstrap

Use the attach file `Verizon.csv` to answer the following question.

Or visit: http://sites.google.com/site/chiharahesterberg/data2/Verizon.csv to import the data.

Verizon is the primary local telephone company (incumbent local exchange carrier (ILEC)) for a large area of the Eastern United States. As such, it is responsible for providing repair service for the customers of other telephone companies known as competing local exchange carriers (CLECs) in this region. Verizon is subject to fines if the repair time (the time it takes to fix a problem) for CLEC customers are substantially worse than those for Verizon customers.

The data set Verizon contains a sample of repair time for 1664 ILEC and 23 CLEC customers. The mean repair times are 8.4 hours for ILEC customers and 16.5 hours for CLEC customers. We're now concerning about **the ratio of two population mean**.

```
Verizon = read.csv("http://sites.google.com/site/chiharahesterberg/data2/Verizon.csv")
#write.csv(x = Verizon, file = 'Verizon.csv')

Time.ILEC = subset(Verizon, select = Time, Group == "ILEC", drop = T)
Time.CLEC = subset(Verizon, select = Time, Group == "CLEC", drop = T)
```

1. If what we concern is the ratio of two population mean, say $\frac{\mu_{ILEC}}{\mu_{CLEC}}$ where μ_i represents the average repair time for customer i, what is the point estimator for it? Use analogy principle.

ANS: Let $\bar{x}_i$ denotes the observed mean repair time for customer i. Then the point estimator by analogy principle is $\frac{\bar{X}_{ILEC}}{\bar{X}_{CLEC}}$.

2. Compute the statistic of interest. In this case, our statistic of interest is related to previous question.

ANS: The point estimator is $\frac{\bar{X}_{ILEC}}{\bar{X}_{CLEC}}$. Then the statistic of interest is its realization: $\frac{\bar{x}_{ILEC}}{\bar{x}_{CLEC}} = 0.5095126$ given the random sample.

```
mean(Time.ILEC)/mean(Time.CLEC)
```

```
## [1] 0.5095126
```

3. Do the bootstrap in R for 10^4 times. (i.e. $B = 10^4$). Remember you should draw the samples with replacement. What is the standard error of the point estimator in (1.)?

ANS

```
B = 10^4
time.ratio.mean = numeric(B)
for(i in 1:B){
  ILEC.sample = sample(Time.ILEC, 1664, replace = TRUE)
  CLEC.sample = sample(Time.CLEC, 23, replace = TRUE)
  time.ratio.mean[i] = mean(ILEC.sample)/mean(CLEC.sample)
```

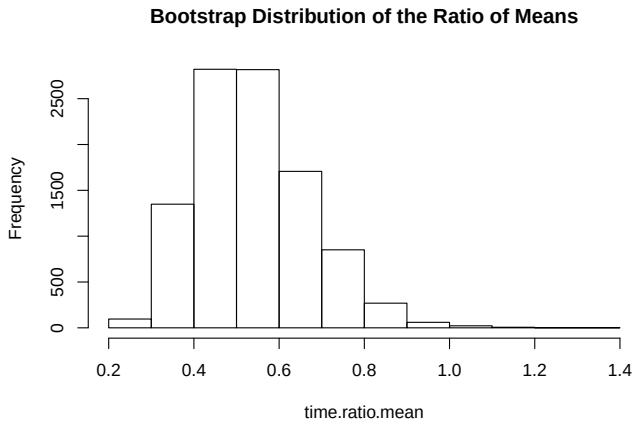
```
}
sd(time.ratio.mean)
```

```
## [1] 0.132785
```

4. Draw the bootstrap distribution with `plot()`function. Is the distribution skew or not?

ANS The histogram shows that the bootstrap distribution of the statistic of interest is skew.

```
hist(time.ratio.mean, main = 'Bootstrap Distribution of the Ratio of Means')
```



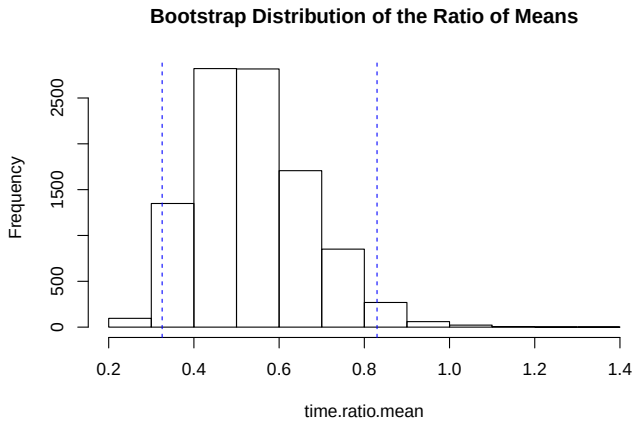
5. Construct the bootstrap interval estimate. Add the upper bound and lower bound onto the previous graph. Use `abline()` function to do so.

ANS

```
#The Interval Estimate with 95% Confidence
quantile(time.ratio.mean, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.3254187 0.8298302
```

```
L = quantile(time.ratio.mean, 0.025)
U = quantile(time.ratio.mean, 0.975)
hist(time.ratio.mean, main="Bootstrap Distribution of the Ratio of Means")
abline(v=L, col = "blue", lty = 2)
abline(v=U, col = "blue", lty = 2)
```



6. If the null hypothesis and alternative hypothesis are the following:

$$H_0 : \text{The repair time for ILEC is half of the repair time for CLEC}$$

$$H_a : \text{The repair time for ILEC is not half of the repair time for CLEC}$$

What will you conclude with a significance level $\alpha = 0.05$? Use the interval estimate you get previously to answer this question.

ANS The interval estimate with 95% confidence in **(6.)** does include $\frac{\mu_{ILEC}}{\mu_{CLEC}} = 0.5$. It tells that we do not reject the null hypothesis that the repair timefor ILEC is half of the repair time for CLEC.

7. If the null hypothesis and alternative hypothesis are the following:

$$H_0 : \text{The repair time for ILEC is equal to the repair time for CLEC}$$

$$H_a : \text{The repair time for ILEC is less than the repair time for CLEC}$$

What will you conclude with a significance level $\alpha = 0.05$? Note that this is a single tail test, you may need to construct a different interval estimate.

ANS Since the test is a single tail test, we need to construct a different interval estimate. Given significance level $\alpha = 0.05$, we want to find a constant c s.t. $Pr(\frac{\mu_{ILEC}}{\mu_{CLEC}} \leq c) = 1 - \alpha$

With bootstrap distribution, we can find the above interval estimate by finding the 0.95 quantile.

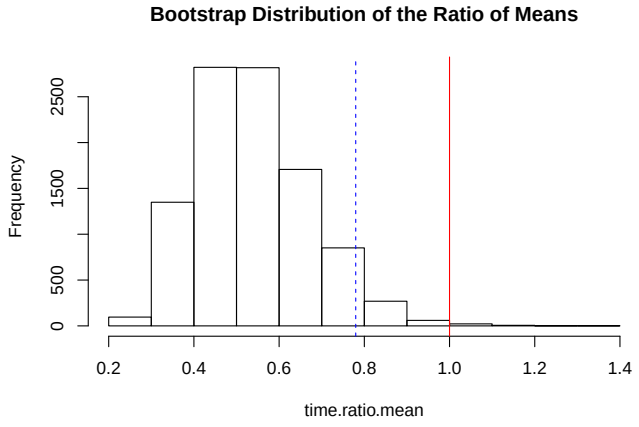
```
#Upper bound
SingleTailUpperBound = quantile(time.ratio.mean, 0.95)
SingleTailUpperBound
```

```
##       95%
## 0.7798465
```

Thus our new interval estimate is: $[0, 0.7734502]$, where 0 is a natural lower bound for this interval since repair time is non-negative. Otherwise the interval should be $\frac{\mu_{ILEC}}{\mu_{CLEC}} \leq 0.7734502$

With the same approach, we can find out that the parameter $\frac{\mu_{ILEC}}{\mu_{CLEC}} = 0$ under H_0 is not included in the above interval estimate. Thus, with the rejection rule, we reject H_0. It means that we reject that the repair time for ILEC is equal to the repair time for CLEC with 5% significance level.

```
hist(time.ratio.mean, main="Bootstrap Distribution of the Ratio of Means")
#The upper bound of the interval estimate
abline(v=SingleTailUpperBound, col = "blue", lty = 2)
#The observed mean ratio
abline(v=1, col = "red", lty = 1)
```



B. Permutation Test

1. Do the permutation test under the following null hypothesis and alternative hypothesis with $\alpha = 0.05$ and $10^4 - 1$ times. Remember that you should draw the samples without replacement.

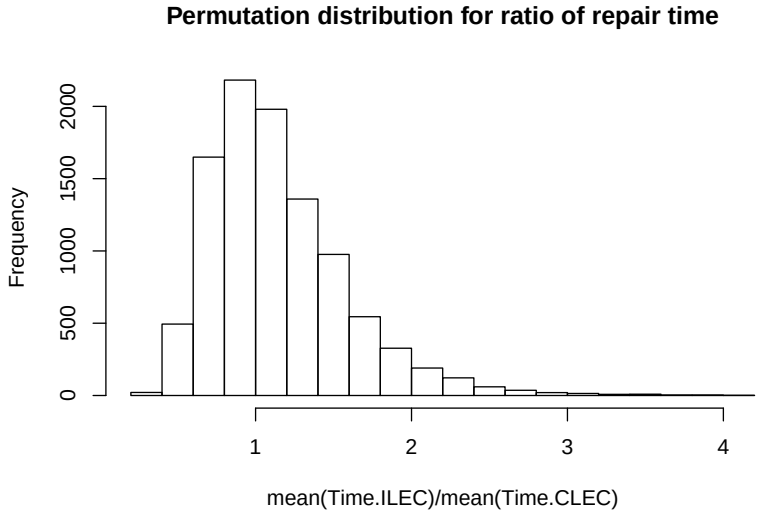
$$H_0 : \text{The repair time for ILEC is equal to the repair time for CLEC}$$

$$H_a : \text{The repair time for ILEC is less than the repair time for CLEC}$$

Draw the distribution of the permutation resamples. Note that the statistic of interest is the same.

ANS

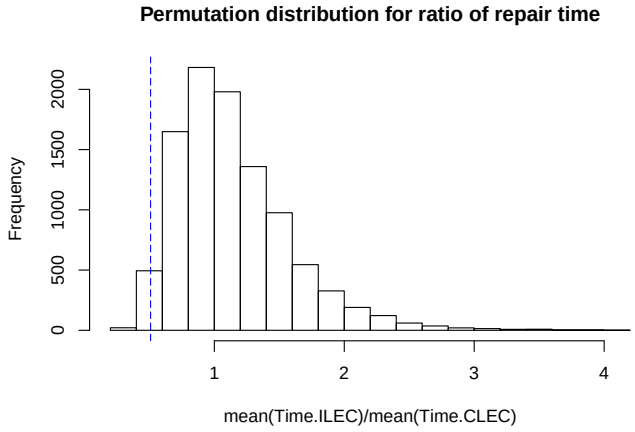
```
repairTime = Verizon$Time
observed = mean(Time.ILEC)/mean(Time.CLEC)

N = 10000-1  #set number of times to repeat this process
result = numeric(N) # space to save the random differences
for(i in 1:N){
  index = sample(1687, size=1664, replace = FALSE) # sample of numbers from 1:1687
  result[i] = mean(repairTime[index])/mean(repairTime[-index])
}

hist(result, xlab = "mean(Time.ILEC)/mean(Time.CLEC)", main = "Permutation distribution for ratio of rep
```



2. Add the observed statistic of interest onto the graph as a vertical line by using `abline()` function

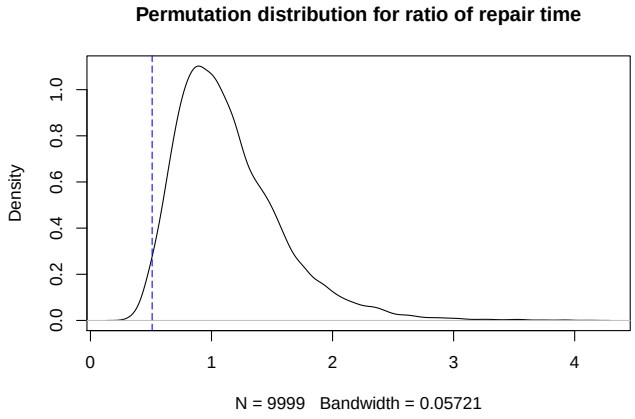
ANS

```
hist(result, xlab = "mean(Time.ILEC)/mean(Time.CLEC)", main = "Permutation distribution for ratio of rep
abline(v = observed, col = "blue", lty=5)
```



Or alternatively:

```
plot(density(result), main = "Permutation distribution for ratio of repair time")
abline(v = observed, col = "blue", lty=5)
```



3. Calculate the p-value. Tell the relation between p-value and the graph in (1.) & vertical line in (2.).

ANS

```
#P-value
(sum(result <= observed)+1)/(N+ 1)
```

```
## [1] 0.0185
```

Under the assumption that the null hypothesis is true, we construct a distribution which is so called permutation distribution. We define the more extreme events by the events that the statistic of interet is less than the statistic observed. We can find the probability with permutation distribution.

The vertical line marks the observed statistic, which is the ratio of sample means (i.e. $\frac{\bar{x}_{ILEC}}{\bar{x}_{CLEC}} = 0.5095126$) The area of the left hand side of this vertical line represents the probability of more extreme events under null hypothesis, which is so called p-value.

With the rejection rule: If p-value is less than the significance level we choose, then we reject the null hypothesis.