Estimation of True Difference in Internet Latency Between Charter Internet and Kalamazoo College ISP

Daniel Michelin, Erik Hartig, and Nicholas Swain

June 11, 2018

Abstract

In this paper we estimate the true μ of the difference between ping times between Kalamazoo College and a residence within a mile located at 416 Oak St. We hypothesize that a significant difference in latency as measured by ping times is likely indicative of substantially different network architecture.

1 Background

A good network connection can be evaluated in a number of ways varying from measuring the amount of data over a network

2 Supporting Statistical Theory

3 Methods

We collected four datasets, two to estimate population parameters and two for exploratory data analysis. Samples collected from Kalamazoo College and the Oak Street residence that are used to estimate the μ of the population of the difference in means were collected at approximately the same time to control for overall network business. Laptops tested used the same wifi chip to account for possible increase in latency there. Latency from router distance was controlled by being approximately the same distance from each one. Latency derived from the router itself is negligable.

4 Results

The first thing we wanted to look at was the difference between the ping time over the course of 180 minutes at both Kalamazoo College and our personal

residence off campus. We collected ping data every 5 minutes at each location. The first thing we wanted to examine was a summary of the data to get a general idea of what was going on.

```
> ping_times <- read.csv("times.csv")
> normality <- read.csv("NormalityCheckTimes.csv")
> normality2 <- read.csv("NormalityCheckTimes2.csv")
> normalitySchool <- read.csv("NormalityCheckTimesSchool.csv")
```

Kalamazoo College Data

```
> KalamazooCollege = subset(ping_times,select=TIME,LOC=="KC",drop=T)
> summary(KalamazooCollege)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00    9.75   15.00   21.75   32.25   57.00
```

Personal Residence Data

```
> PersonalResidence = subset(ping_times,select=TIME,LOC=="OK",drop=T)
> summary(PersonalResidence)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   24.00   27.00   29.31   31.25   67.00
```

Based on this data alone we can see that in general it appears that Kalamazoo College appears to be generally faster than the personal residence. However we wanted to see if this difference in mean ping time was statistically significant. We decided to use a permutation test to do this.

5 Calculation

6 Conclusions

```
> # P value
> obs = mean(KalamazooCollege) - mean(PersonalResidence)
> N <- 10^4
> result <- numeric(N)
> for(i in 1:N){
+   index <- sample(72,36,FALSE)
+   result[i] <- mean(ping_times$TIME[index]) - mean(ping_times$TIME[-index])
+ }
> # 2 sided test p value
> ((sum(result < obs)+1)/N)*2

[1] 0.0098
```

Since our value is less than .05 it is safe to say that the difference in means is statistically signifigant. Now that we know it is signifigant we wanted to find a confidence interval to see what kind of range we could expect. We used bootstrap to find this interval.

```
> N <- 10^4
> times.diff.mean <- numeric(N)
> for(i in 1:N){
+   x <- sample(KalamazooCollege, 36, replace=TRUE)
+   y <- sample(PersonalResidence,36, replace=TRUE)
+   times.diff.mean[i] <- mean(x) - mean(y)
+ }
> hist(times.diff.mean, main = "Bootstrap distribution of difference in means")
> abline(v = mean(KalamazooCollege) - mean(PersonalResidence), col = "blue", lty = 2)
> dev.new() # Open new graphics device
> qqnorm(times.diff.mean)
> qqline(times.diff.mean)
> obs = mean(KalamazooCollege) - mean(PersonalResidence)
> mean(times.diff.mean)

[1] -7.569208

> obs

[1] -7.555556

> quantile(times.diff.mean,c(0.025,0.975))

      2.5%      97.5%
-13.277778  -1.888889
```

So this means we can be pretty sure that the school has a ping time that is between 2 and 13 ms faster than our personal network. With that being known we wanted to examine the kind of distribution that the data had. With that in mind we than examined a histogram of both of the data sets to try and figure out what kind of distribution we have.

```
> hist(KalamazooCollege)
> hist(PersonalResidence)
```

Based on this we were not really sure what kind of distribution these were. Based on the histograms we thought that the personal residence data might be normally distributed so we wanted to test that using qqnorm.

```
> qqnorm(PersonalResidence)
> qqline(PersonalResidence)
```

Based on this we thought this looked fairly normally distributed with a few outliers or some right skew. Therefore we wanted to see how close the Kalamazoo College ping time fit the normal distribution.

```
> qqnorm(KalamazooCollege)
> qqline(KalamazooCollege)
```

This did not appear to really follow a normal distribution. ALthough parts of it fell on the line it seems to vary from it quite signifigantly. After this realization we wanted to test if ping followed a normal distribution as a rule of thumb. So we ran a test on our personal network that recursivly pinged google every 2 seconds 200 times. Than we examined the histogram created by that data.

```
> hist(normality$TIME,breaks=50)

> #normality check
> hist(normality$TIME,breaks=50)
> qqnorm(normality$TIME)
> qqline(normality$TIME)
> barplot(normality$TIME)
> #normality check inside of a "chunk"
> hist(normality2$TIME[111:190],breaks=50)
> qqnorm(normality2$TIME[111:190])
> qqline(normality2$TIME[111:190])
> barplot(normality2$TIME[111:190])
> boxplot(normality2$TIME[111:190])
> summary(normality2$TIME[111:190])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.00   26.00   31.00   29.81   33.00   42.00

> #normality check at school
> hist(normalitySchool$TIME,breaks=50)
> qqnorm(normalitySchool$TIME)
> qqline(normalitySchool$TIME)
> barplot(normalitySchool$TIME)
> # Summary statistics
> summary(ping_times)

 LOC          TIME
 KC:36   Min.   : 6.00
 OK:36   1st Qu.:15.50
         Median :26.00
         Mean   :25.53
         3rd Qu.:32.00
         Max.   :67.00

> hist(ping_times$TIME)
> KalamazooCollege = subset(ping_times,select=TIME,LOC=="KC",drop=T)
> PersonalResidence = subset(ping_times,select=TIME,LOC=="OK",drop=T)
> hist(KalamazooCollege)
```

```
> hist(PersonalResidence)
> qqnorm(KalamazooCollege)
> qqline(KalamazooCollege)
> barplot(KalamazooCollege)
> barplot(PersonalResidence)
> var(KalamazooCollege)

[1] 221.1071

> var(KalamazooCollege)

[1] 221.1071

> mean(PersonalResidence)

[1] 29.30556

> mean(KalamazooCollege)

[1] 21.75

> #Theoretical confidence interval
> t.test(KalamazooCollege,PersonalResidence)

        Welch Two Sample t-test

data:  KalamazooCollege and PersonalResidence
t = -2.5939, df = 58.308, p-value = 0.01198
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.385513  -1.725598
sample estimates:
mean of x mean of y
 21.75000  29.30556

>
>
>
>
```