Analysis of Bandwidth in the United States

Ali Taheri, Mike Hendrickson

Department of Data Science, School of Engineering and Computer Science, University of the Pacific
{s_taheritari , m_hendrickson}@u.pacific.edu

***Abstract*—In this study the researchers used the built in statistical tools in R such as t-tests, box plots, bar graphs and linear regression models to test the validity of the average upload and download bandwidth speed claims made by the company OOKLA© as well as show the relationship between the number of people per household and purchased bandwidth speed. With the Internet as the most direct line to providing information for households, businesses, schools and other industries, looking to provide better, faster and unfettered access to information is very important these days. With the vote to end Net Neutrality in the coming weeks (December 14, 2017), allowing internet service providers to control not only how fast people receive information, but what information they could receive, we look to understand the general current state of bandwidth availability and contemplate the upcoming changes.**

I. Introduction

Nowadays, The Internet is as important as food, water and even oxygen for a lot of people. About 15 years ago it was a luxury service and life without Internet was possible. It is interesting that one major factor of development of countries is the percentage of people with access to Internet and moreover speed of their connection. Consumers of the Internet are required to have access to quick and unfettered information for business, educational, and entertainment purposes. Whether a consumer is accessing the Internet from home, school, work or the library, it has become a utility rather than a privilege in our everyday lives. In order to stay ahead in the capacities listed above, consumers purchase as much bandwidth speed as they can afford and they believe to be necessary. One important parameter of internet connections is its speed. At first, average speed of connections was about 33.6 Kbps (Kilo bits per second) but with exponential progress in technology the capabilities of bandwidth are now 1 or even 10 Gbps(Giga bits per second). A lot of people are "bandwidth-hungry" these days and telco operators are doing their best to offer ultra high speed services to their clients. Additionally, there is a need for not only fast, but reliable connections to the Internet for the purposes mentioned above and, in more recent days, IoT or Internet of Things services. Based on these new requirements, in 2014, the average consumer held 2.9 connected devices in order to stay relevant in the world of information [1].

The researchers conducted a survey to conclude whether or not a previous study made by Speedtest.com from OOKLA correctly predicted the average bandwidth per household and to see if they can produce a simple linear regression model to predict how much bandwidth is being purchased per household based on the number of occupants in the residence. Access to the internet is broken into two different speeds, download speed and upload speed. For residential customers, which this study is focused on, download speeds are higher than upload speeds due to the asymmetric nature of residential access. Because residential consumers need to download information from the internet more than upload information to the internet, an asymmetric access is provided to keep costs down. It is clear that SME (Small to Medium Enterprises) users and big companies have a lot of traffic between their branches, so their traffic must be symmetric.

Speedtest.com, a website that measures bandwidth speeds globally, conducted a study stating that the average residential download speed per residence is 70.75 Mbps (Megabits per second) while the average residential upload speed per residence is 27.64 Mbps in US [2]. For more clarification, download speed that people would see in some applications same to IDM©(Internet Download Manager) is based on megabytes per second, so it must be multiplied by 8 to be as megabits per second.

The researchers used these speeds as the null hypotheses for this study.

For the hypothesis test of download speed: (μ is average speed for all of US residential users)

$$\begin{aligned} H_0 &: \mu_{download} = 70.75\,Mbps \\ H_a &: \mu_{download} \neq 70.75\,Mbps \end{aligned} \tag{1}$$

For the hypothesis test of upload speed:

$$\begin{aligned} H_0 &: \mu_{upload} = 27.64\,Mbps \\ H_a &: \mu_{upload} \neq 27.64\,Mbps \end{aligned} \tag{2}$$

II. Dataset

The researchers collected the data for this study from a six question survey created on Surveymonkey.com and distributed them to friends, family, colleagues and classmates via email, phone calls and word of mouth in the fashion of convenience sampling. In order to diversify the sample data geographically, the researchers specifically sent the survey to their known associates in states outside of their home state, California. The survey stayed open on Survey Monkey between November 21st, 2017 and November 28th, 2017. The questions on the survey were as following:

- In which state do you live?
- How many people use the internet in your house?
- Who is your Internet service provider?
- What is your download speed?
- What is your upload speed?
- What form of media connects you to the Internet?

The researchers received 37 total responses. Household size of responders ranged from one to six persons per each. Comcast, Verizon and AT&T were among the telco provider responses. Download speed responses ranged between 10 and 150 Mbps and upload speed responses ranged between 1 and 55 Mbps. Alabama, California, Georgia and Tennessee were among the geographic responses. Cable, fiber and wireless interfaces made up the responses for media type. It is obvious that high speeds are attainable mainly on fiber optics and other media would provide these speeds in special circumstances same to very short distances and low noise environments .

A. Univariate Survey Response Analysis

For the hypothesis tests, the researchers were interested with the download and upload speed variables. During analysis, the researchers removed samples from the dataset that did not include answers for either download or upload speeds resulting in a total of 34 responses for download speed and 29 responses for upload speed. The dataset provided an average download speed of 71.47 Mbps and an average upload speed of 16.21 Mbps.

B. Jointly Distributed Survey Response Analysis

The researchers used the same dataset for the jointly distributed survey response analysis as was used for the univariate survey response analysis. They continued to use the results with the null values removed. In order to try to create a simple linear regression model to predict the download speed of a household based on the number of people living in the residence, the number of people per household was used as the independent variable using a ratio scale of measurement and download speed was used as the dependent variable using a ratio scale of measurement too.

III. Methodology

Survey responses were downloaded directly from the Survey Monkey website to an .xlsx file and uploaded to R for analysis. The libraries of stats, knitr, dplyr, kableExtra and gdata of R were called for statistical and graphical analysis. Bar plots were created for the initial exploratory analysis to view the distribution of average download speed per state, average download speed per number of people in the household, average download speed per Internet service provider, samples per form of media and a box plot was created to show the gathered data about speed.

A. Univariate Survey Response Analysis

A t-test was created for both the download speed hypothesis test and the upload speed hypothesis test using the built in t.test function in R because the population standard deviations of both download and upload speeds were unknown to the researchers. The test used a 0.01 significance level to provide a more confident answer.

B. Jointly Distributed Survey Response Analysis

In order to create a simple linear regression model, the researchers compared the number of people per household to download speed responses using the built in R function of "lm". The researchers used a scatter plot to plot the responses and plotted the linear function created to have an initial view of the fit. Though the linear function did not appear to fit the samples well visually, the researchers continued to check the validity of the function using the LINE assumptions: linearity, independence, normality of residuals and equal variance (homoscedasticity). Linearity and homoscedasticity were tested through a scatter plot of the residuals vs. the number of people per household and a plotted horizontal line on the origin. Because the scatter plot showed residual variance decreased as the magnitude of number of people per household increased, the researchers tried a boxcox transformation on the data, but it was not successful in creating more equal variance. To check for normality, the researchers created a Q-Q Plot using the built in qqnorm and qqline functions in R.

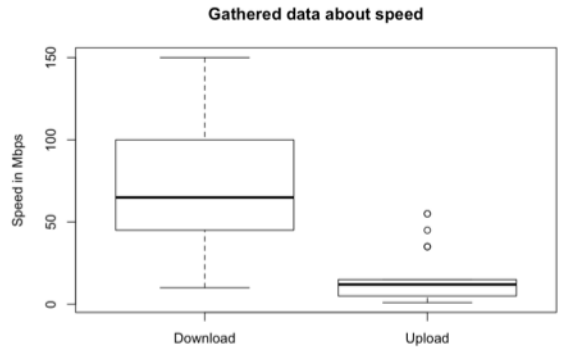
IV. Results and Discussion

Though the linear assumptions were not met and the sample size was small, some important insights can be reached from the researchers results.

A. Univariate Survey Response Analysis

From the t-test determining whether or not the average mean of download speeds per household was 70.75 Mbps as recorded by Speedtest, the researchers found a p-value of 0.9044. In addition to the 71.47 Mbps average of sample, a p-value of that magnitude provides overwhelming evidence that researchers do no have sufficient evidence to refuse Speedtest's claim about the average download speed per household.

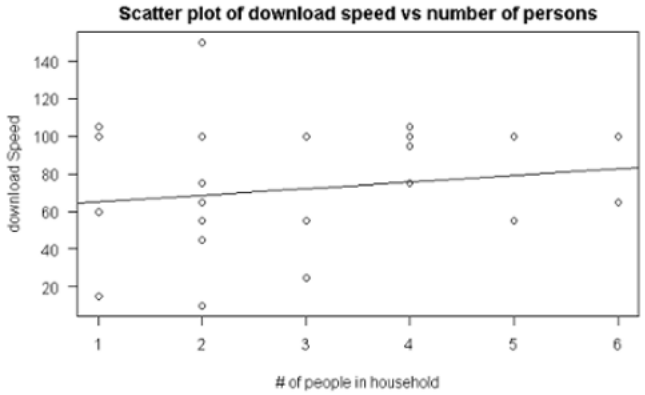
From the t-test determining whether or not the average mean of upload speed per household was 27.64 megabits per second as recorded by Speedtest, the researchers found a p-value of 0.0004. From the earlier discussion explaining that due the asymmetric nature of the residential bandwidth speeds, it is reasonable to assume that the average upload speed is going to be lower than 70.75 megabits per second. However, with such a small p-value, at a significance level of 0.01, the researchers found sufficient evidence to reject Speedtest's claim that the average household upload speed in the United States is 27.64 Mbps.



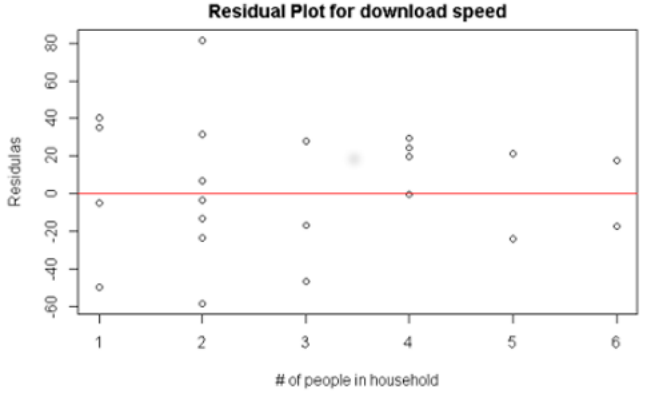
B. Jointly Distributed Survey Response Analysis

The simple linear regression model's formula created using the sample data is:

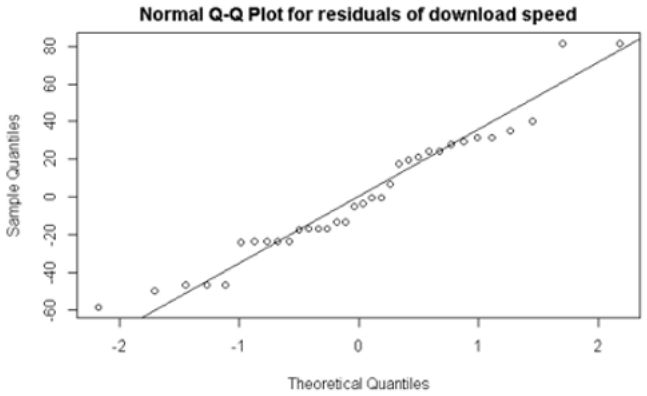
$$download_speed = 61.418 + 3.524(persons) \quad (3)$$



The residual vs. number of people per household plot, as discussed earlier, shows that a linear model is a moderate fit, but that there is an obvious decline in variance as the magnitude of people per household increases. Logically, this makes sense as cable internet products typically have a limit of around 100 megabits per second of download speed, so larger families will more often purchase as much bandwidth as possible. At the same time, smaller households have the option to purchase less bandwidth as they see fit or purchase more than necessary. Because of this, the homoscedasticity assumption is not met.



The Q-Q plot shows the normality of residuals assumption can be moderately accepted. There are two potential outliers sitting outside of the -2 and 2 quartiles, but they were not taken out of the study as they are very close to the 2 and -2 quartiles.



The results of the scatter plot and linear function show, at best, a weak relationship, and with an R^2 value of 0.019, the researchers can only attribute 1.9 percent of the variance in bandwidth to the number of people in a household. This is indicative that consumers do not know how much bandwidth they need, and with marketing ploys and general availability, may be paying for a lot more bandwidth than they need. The average downloading activity for each person including emails, streaming, gaming, etc. requires no more than 5 to 10 Mbps. This means that a house of seven simultaneous users are required in order to utilize a full 70 megabit per second connection assuming they are each performing one average downloading function. From the scatter plot, households of six are averaging around 90 megabits per second, which in this multitasking era seems reasonable and maybe even low, households of two are paying for the same speed and more.

Due to the researchers industry knowledge, it was concluded that the 150 megabit per second responses were from fiber media. Fibers carry optical signals and attenuation of them is very lower than electrical signals on copper and cable, so their speed would be very high. Fiber connections have not become the mainstream for residences in the United States, but are available in some larger metropolitan areas.

V. Conclusion

The researchers considered a number of limitations to this research. Because of a time constraint, the sample size was small and the researchers used a survey to receive as many responses as possible by reaching out to friends, family, coworkers and classmates. Studies using this type of convenience sampling tend to have less external validity. Because both researchers are in the telecom field and studying data science, the friends and families will tend to be of like mindedness and those in scientific fields tend to buy as much bandwidth as possible for multitasking. In addition, because over half of the responses were received from those living in or around the San Francisco area, a metropolis with

a multitude of internet availability, a large portion of the United States was not considered in this study. With middle America having typically larger families and lower bandwidth availability, that could have change the results dramatically. The second largest issue is internally within the survey. Most are familiar with only the advertised speed of their internet connection. However, real bandwidth speeds can be much higher or much lower than advertised. To create a better study, an experiment could be completed by connection to random locations around the United States and using a product to measure real bandwidth speeds. Because the null hypothesis was based on speedtest.com, a different product could be used to measure or even a few products randomly assigned to random locations. After the experiment, the researchers would run the same linear regression model to verify the relationship between number of people per household and bandwidth. If the results did not change, the conversation starts with whether or not consumers are paying too much for their internet connection. The study could also be taken outside of the United States to explore Internet availability around the world. With that data, it could be possible to begin understanding how access to the internet affects development of countries.

While the researchers were unable to find a good predictive model for measuring bandwidth using household size, they were able to deduce other possible reasons for purchased bandwidth including marketing strategies and general availability. Along with that discovery came a philosophical question to consider: Should households in less developed areas not be allowed the same access to the internet, a direct line to information, as developed cities? How would that affect business, growth and education?

REFERENCES

[1] 'Average number of connected devices used per person in selected countries in 2014', 2016. [Online]. Available: https://www.statista.com/statistics/333861/connected-devices-per-person-in-selected-countries/.

[2] K. Murnane, 'Speedtest Ranks Internet Access Speed In More Than 100 Countries', 2017. [Online]. Available: https://www.forbes.com/sites/kevinmurnane/2017/08/14/speedtest-ranks-internet-access-speed-in-more-than-100-countries/.