STA238 - Winter 2021

Assignment 3

GROUP 194: Leuven Wang

March 5, 2021

Introduction

As of 2019, the TTC operates 163 bus routes across Toronto, carrying over 235 million passengers annually (Toronto Transit Commission, 2019). Being an integral part of Toronto's public infrastructure, it is of great economic importance that the bus service should be as effective and efficient as possible. One measurement we can use to understand its efficiency better is the duration of bus delays. This is the amount of time between a bus' scheduled arrival time at a stop and the time it actually arrived when the bus is behind schedule.

This investigation will aim to answer two main questions. Firstly, what is the true mean duration of bus delays across all TTC bus trips that are late? Secondly, we will measure the total duration of delays across all incidents and see what proportion of the delay times can be attributed to delays that occurred during rush hour. For the purposes of this investigation, we will treat rush hour as the hours between 8-9am and 5-6pm on weekdays.

Our data in this case is the TTC's record of delays for each incident where a bus is late from January to August 2020. It includes the date, time, and duration of each delay. We will assume that this data is proportionally representative of all TTC bus delays across an extended period of time.

Our hypotheses are that the true mean duration of all bus delays will lie somewhere between 10 to 20 minutes and that the proportion of delay time that occurred during rush hour is greater than $\frac{1}{12}$ and smaller than $\frac{1}{9}$.

Data

Our data is provided by the TTC and it records every instance between January and August 2020 where a bus was behind schedule i.e. late. It documents the duration of the delay, the time at which it actually arrived at the stop, and the day. It does not store cases where a bus is early or on time. Thus, we are only looking at cases where a bus is late.

Data Cleaning

Most of the data was comprehensive and complete. Nevertheless, we removed any records with missing values and chose only to retain the variables that are necessary to our investigation i.e. the duration of the delay and the day and time. We also converted the time into a numerical format so that we can sort it using maths later on. The important variables of our investigation are:

Data Description

Delay - The time between the bus' scheduled arrival time and its actual arrival time at a station in minutes.
Day - The day of the week in which this lateness occurred e.g. Sunday.
Time - The time of the day at which the bus arrived late.

Summary	**Measurement**
Mean Delay Time in Minutes	19.7601396
Median Delay Time in Minutes	10
Standard Deviation of Delay Time in Minutes	68.366731
Proportion of Delay Time Attributable to Rush Hour	0.079046

Fig 1: Table of Numerical Summaries of Sample Data

Figure 1 illustrates some conceptions of the center and spread of the duration of delays. This helps us understand how long most delays are and how much we can expect the duration of delays to vary. It also shows the proportion of delay time attributable to delays during rush hour to give us an idea of what we may expect later on.

Fig 2: Density Histogram of Delay Times less than 50 Minutes

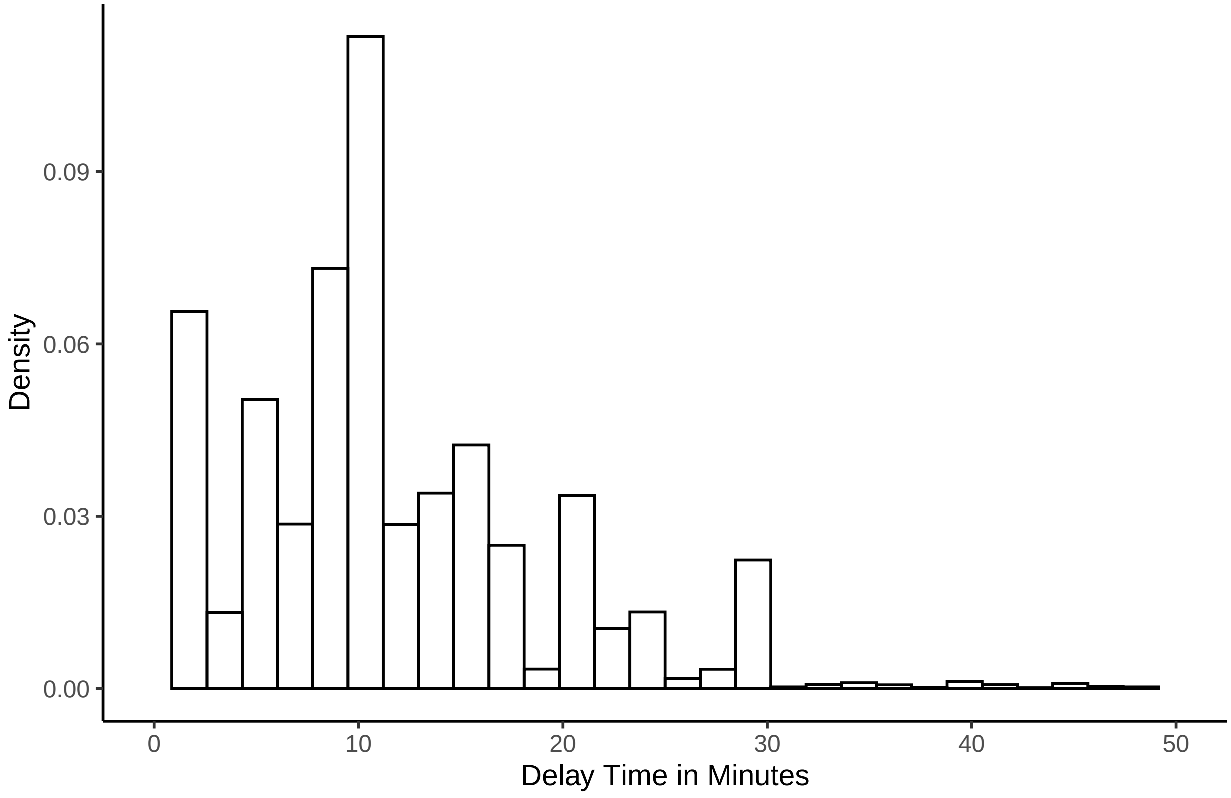


Figure 2 shows the density histogram of instances where the delay of a bus was less than 50 minutes. The limit of 50 minutes was chosen to make the graph more aesthetically pleasing but it nonetheless demonstrates most of the density distribution we are trying to investigate. There are two main distinctive peaks here - one when the delay time slightly exceeds 10 minutes and one when it is less than 2 minutes.

All analysis for this report was programmed using `R version 4.0.2`.

Methods

To understand the behavior of TTC bus delays across all instances where they are late, we will assume that the data we have is proportionally representative of them and we will project their results with bootstrapping. The specific method we have chosen to proceed with is empirical bootstrapping.

Bootstrapping is a technique used to approximate sample estimates where we do not possess data on an entire population and instead must rely on 1 representative sample to extrapolate conclusions. In bootstrapping, we take observations from that one original sample, allowing for replacements, to create an entirely new sample. We then calculate, from this bootstrap sample, whatever sample statistic we are interested in measuring. We repeat this process to build a large number of bootstrap samples each with its own sample statistic. This creates a sampling distribution - a distribution that includes all the values of the sample statistics we have measured over many bootstrap samples.

We have specifically chosen the empirical bootstrap method as we have no idea what kind of distribution the original data follows or what distribution the entire population of bus delay duration follows. This means that we will be building our bootstrap samples directly from the empirical data in our data set, not attempting to estimate any parameters for any distributions as we have no strong indication of either.

With this technique, we hope to create two sampling distributions - one for each of our measurements: the mean of the duration of all delays and the proportion of delay time caused by delays during rush hour. With these sampling distributions, we will set confidence intervals whereupon we can speak to what degree of confidence we can expect the true value of these measurements to lie within these intervals. Our confidence interval will be set to the middle 95% of our bootstrap distribution such that both ends of the distribution not included in the interval will each equal 2.5% of the distribution. For this investigation, we will be building 1000 bootstrap samples for each metric to create our sampling distributions.

The mean duration of all delays is the mean length of time between a bus' scheduled arrival time and its actual arrival time for all cases where the bus is late.
The proportion of delay time caused by delays during rush hour is a percentage whereupon we measure the total time elapsed by all delays altogether and then use it to divide only the total time elapsed by all delays that occur during rush hour.

To measure these metrics we are mainly using the Delay variable which is a numeric value. The Day and Time variables are only used to identify those observations of delays that occurred during rush hour. Day is categorical whilst we have converted Time to a numeric value so that we can isolate the rush hours using maths.

All analysis for this report was programmed using `R version 4.0.2`.

Results

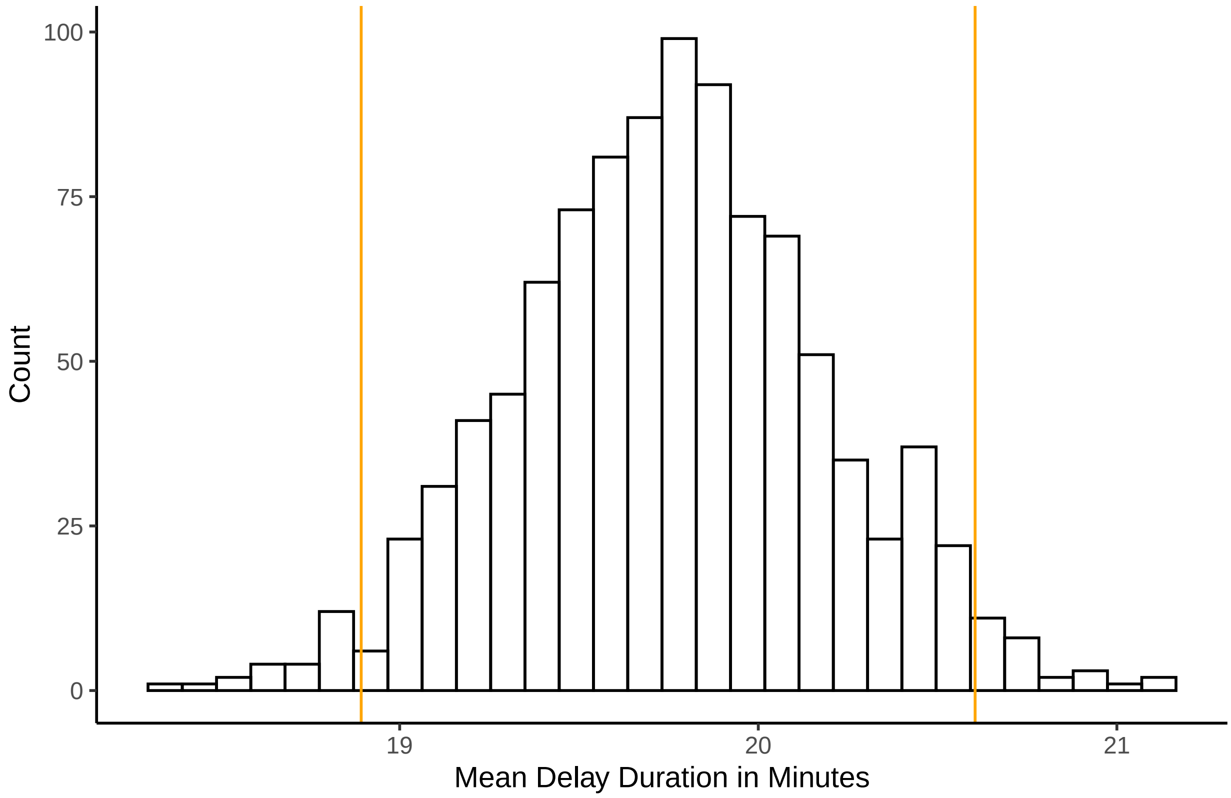
Sampling Distribution	**2.5th Percentile**	**97.5th Percentile**
Mean Delay Time (Minutes)	18.8926806	20.6046446
Proportion of Delay Time Attributable to Rush Hour	0.06885	0.091796

Fig 3: Table of Percentiles for Sampling Distributions

Figure 3 is a table showing the 2.5th and 97.5 percentile for both sampling distributions we have created. This is in line with our plan to create a confidence interval encapsulating the middle 95% of the sampling distributions. What this means is that we are 95% confident that the true mean delay time is between 18.8926806 and 20.6046446 minutes. If we were to repeat this experiment a large number of times, each time creating a sampling distribution and setting a 95% confidence interval, then 95% of those intervals would contain the true mean delay time. Similarly, we can say that we are 95% confident that delays during rush hour account for between 6.8850037% and 9.1795981% of the total time delayed by all late buses.

Both distributions, along with the limits of this confidence interval, are graphically represented in Figure 4 and Figure 5.

Fig. 4: Bootstrap Sampling Distribution of Mean Delay Time



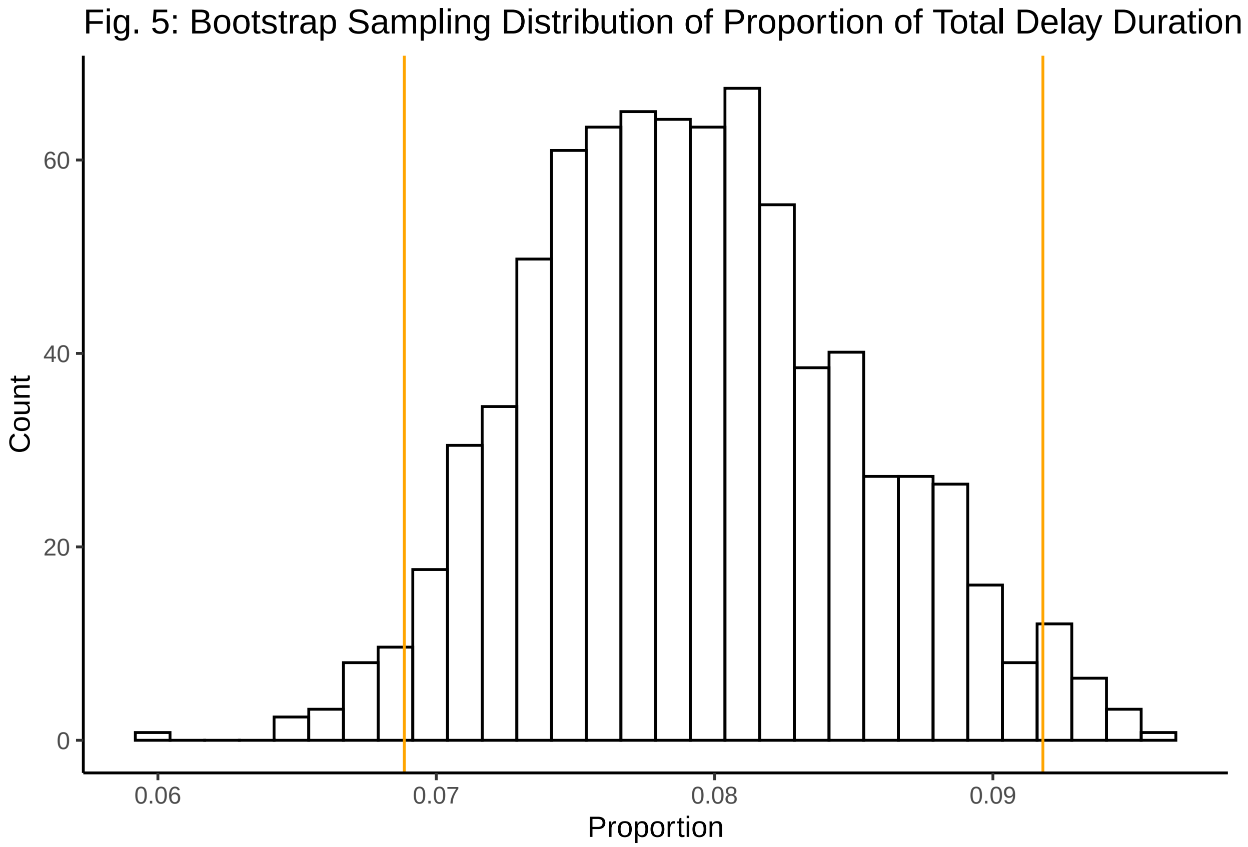


Figure 4 shows the sampling distribution of the mean delay duration whilst Figure 5 shows the sampling distribution of the proportion of total delay times attributable to delays during rush hour. The two orange lines in each graph represent the 2.5th and 96.5th percentiles - all the bars within these two limits are considered part of the confidence interval. That is, we are 95% confident that the true measurement of each metric exists within it. Both distributions somewhat resemble a normal distribution with their peaks in the center.

Our primary assumption for this investigation was that the months from such January to August 2020 were representative of TTC bus patterns as a whole. From one aspect, the results we have achieved seem reasonable. For one thing, our confidence interval for mean delay time includes the mean we calculated from our original sample. The same goes for the proportion of duration of delays that occurred during rush hour.

However, the results achieved for mean delay time does not align well with the histogram of our original sample of all delay times recorded. There, the peak indicated that most delays only last around 10 minutes, hardly within the 20-minute range suggested by our confidence interval. The reason I suspect for this drastic difference between the mean of the original and of the bootstrap samples versus the median of the original sample is that each sample contains anomalous observations where the delay time is drastically higher for a few instances. This has a much more pronounced effect on the means of all our samples and thereby extended them so that they would not reflect what the average commuter usually experiences in delays.

Moreover, it should be noted that given 2 rush hours in a day and 5 workdays in a week, there are 10 rush hours weekly and 168 hours in each week in total. If we were to divide this equally then each two hour period should be responsible for 5.95% of the total amount of time where buses are delayed. However, this is far below our achieved range here of 6% to 9% for the proportion of duration of delays that occurred during rush hour. This suggests that rush hour is responsible for either more frequent delays or longer delays or both.

Conclusions

In this investigation, we have attempted to approximate the true mean duration of all bus delays and the proportion of delay time attributable to rush hour. Given sample data on bus delays between January and August 2020, we have used the method of empirical bootstrapping to build 1000 bootstrap samples for both metrics by taking observations from the original sample with replacement. We then calculated the statistics (mean and proportion) we were interested in for each bootstrap sample and built a sampling distribution for each statistic. We then set a 95% confidence interval on both sampling distributions.

Recall our hypotheses that the true mean duration of all bus delays will lie somewhere between 10 to 20 minutes and that the proportion of delay time that occurred during rush hour is greater than $\frac{1}{12} = 0.083$ and smaller than $\frac{1}{9} = 0.111$. Our results, set within a 95% confidence interval shows that we are 95% confident that the true mean duration of all bus delays is between 18.9 and 20.6 minutes. For the most part, this interval overlaps with the interval in our hypothesis though there is some exclusivity. This means that it's perfectly possible for the true mean duration to extend outside of the range of our hypotheses. Similarly, our 95% confidence interval for the true proportion of delay time due to delays during rush hour overlaps somewhat with our hypothesis interval. We are 95% confident that the true proportion of delay time is between 0.069 and 0.092. This covers a lower portion of our hypothesis range so it is possible that the true proportion of delay time attributable to rush hour exists outside of our hypothesis range.

Regardless of how well aligned our confidence intervals and hypotheses intervals are, we can still extrapolate several interesting conclusions from our results. As pointed out above in the Results section, the two-hour period of rush hour accounts for more delay time than if all two-hour periods split the total delay time equally between them. This means that rush hour is responsible for either more frequent delays or longer delays or both. This opens up an avenue whereby city officials now know that they should focus on minimizing delays during rush hour as opposed to other hours since rush hour accounts for delay lengths disproportionately. We can thus save resources as we pursue policies that specifically address issues caused by rush hour. Some possible next step for policymakers interested in improving the efficiency of the TTC bus service is to look at which stations experience the longest delays and whether the disproportionate total delay time in rush hour is due to lengths of delays or a frequency of delays. For example, if the problem during rush hour appears to be that buses are getting delayed more frequently, then simply adding more buses to a route during rush hour might not be a good solution.

The vast difference between the median of delays duration in the original sample and the suggested confidence interval of the true mean of delays duration in the bootstrap distribution is indicative that the true mean here may not be the most representative measure of what commuters are most likely to experience. This means that most average commuters would probably not have to, in their daily lives, wait 20 minutes past a bus' scheduled arrival time if it is late as suggested by the confidence interval our simulation has projected. This is because, as mentioned above, our simulation took several anomalously large delays in building its samples. As such, future investigations that target the experience of the common commuter may be more interested in using the median or mode as their measurement of average delay time or consider removing such anomalies. This provides a much more representative and accurate depiction of what daily travelers undergo.

There are some limitations to this investigation. For one thing, our original sample data only covers January to August 2020. This is a period marked by tumultuous lockdown policies due to the Covid-19 pandemic. As such, travel patterns and transportation schedules may be different from what is usually experienced in normal circumstances. For example, the intensity of rush hour may have been lessened in this period as more office workers work from home and do not commute, thereby leading to less crowded roads and fewer delays. One should also investigate if the TTC modified the frequency and scheduling of their buses during this time. Additionally, please note that the period from January to August does not encapsulate the entire year and therefore, even if 2020 was a normal year, we would have to assume that the months from September to December have similar transportation patterns. This would have to take into account weather patterns and so forth. Moreover, this investigation only looked at data where the buses were already late. It does not speak to the average waiting time for a Toronto bus as it does not account for cases where buses were on time or early. Thus it does not tell us anything about how often Toronto buses are late.

In conclusion, our investigation has revealed some interesting insights into the nature of TTC bus delays. We have discovered that rush hour seems to disproportionately account for the length of total bus delays. We have also established that the mean may not be the best metric used for ascertaining what the average commuter would experience in delay times. We have made some recommendations as to what future investigations aimed at tackling efficiency problems in the TTC might involve. This includes a focus on whether longer total delay times during rush hour is attributable to longer delays or more frequent ones. We have noted several underlying factors such as the Covid-19 pandemic that may challenge the assumptions of this investigation although the extent to which this challenge is justified is debatable. We have also noted the drawback of only examining cases where TTC buses are already late. Nevertheless, this investigation has served well in helping us to understand some aspects of the TTC bus service. This is an important step for anyone seeking to improve the efficiency of public infrastructure in Canada's largest city.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. Toronto Transit Commission *TTC 2019 Operating Statistics.* http://www.ttc.ca/About_the_TTC/Operating_Statistics/2019/index.jsp. (Last Accessed: February 27, 2021)
5. Toronto Transit Commission *TTC Bus Delay Data.* https://open.toronto.ca/dataset/ttc-bus-delay-data/.(Last Accessed: February 26, 2021)
6. All analysis for this report was programmed using `R version 4.0.2`. Libraries used: openintro, tidyverse, dplyr, opendatatoronto