Part 1

Goal

As we all know, students go back to have class in person need different methods to commute between home and school. As a U of T student, there are also plenty of choices for each student to choose.(University of Toronto, 2022) Although there are plenty of transportation to choose, students still face the lateness problem. Therefore, it is essential for government and U of T to consider offering different kinds of transportation methods and its quickness, try to reduce the lateness for students. At first, knowing the most popular commuting way for students is essential and this will allow the University of Toronto to focus more on building more commuting options that suit the needs of college students. In my survey, I will investigate which transportation method is the most popular for U of T students to go for school and how often students late for school in year 2022. My purpose is to provide some suggestions for the University of Toronto on how to create a better commute for students and reduce the lateness rate for students in the future. The topic is interesting since students of U of T can go to school more handy with less probability to be late and U of T can have more wonderful facilities about transportation.

Procedure

Since my survey is to investigate U of T students who go for school this year, thus my target population is all U of T students who go for in person school in 2022. In order to get the permission of knowing their commutation methods, I first need to notify U of T my purpose of this survey and after knowing my purpose, I would ask them if they can give me some students' U of T mail about U of T students in order to connect with them. Thus, my frame population is all the students who I can connect through sending email to them. Then the students who receive my emails and response to me will be my sample population. (If the first time they don't response to me, I will send again.) Then, in order to perform by using stratified sampling method, I want ask them their distance to school and classified them into different groups by the distance to school. After collecting all information from different groups, I would randomly choose some students perhaps 20 - 30 students to ask them helping me finish the survey. I think my strengths of my proposed sampling procedure is that different students in different address will perform different. Like, some students live much far away from school, they would prefer driving or taking TTC and so on. By doing stratified sampling, I can get more accurate results. On the opposite, my survey has some flaws that are inevitable. Since I would send email to students, many students will think my email is junk email or fraudulent email that they don't want to response. Also, it is hard to ask for U of T to give me some U of T mail, since university may think they are leaking student information.

Showcasing the survey.

https://docs.google.com/forms/d/1fujpCYhCLjP1HgR7KunQtXwyjC-WFF551bj_Yhl2ZgU/edit[1]

Question 9: The purpose for me to create this question is to collect how often students late for school once a week, this may give suggestions for U of T if it can offer students some possible ways to help them reduce the probability that late for school. The pros are to get information directly and question only includes students who take in person classes. The cons are students may late for school different times a week and students may not want to share the truth for me since late for school is a shame for them.

Question 4: This question aims to ask students their actual commute transportation. Since by recording students actual transportation, U of T can have more accurate data, then it can offer more help for the most popular mode of transportation. Pros: This question considers many students may take different modes of transportation during a week, then a multiple choice is better than a single choice. Cons: Sometime, students may think this question is redundant, since another question ask a similar one which ask their preference, and their preference is just the same as actual. On top of that, some students don't go to school, thus it is not useful for them to answer this question.

Question 5: Question 5 is to get response from students why they choose that mode of transportation. It is reasonable to know the actual reason that students can give a better suggestions for U of T on improving the basic measures for traffic in U of T. Pros: Increase the persuasiveness of my advice to schools. This allows

schools to recognize that more people have to choose a certain commuting method for what reasons, and then the school can better take measures to give students a better commuting experience. Cons: Since this is a short answer, many students don't want to reply seriously and then the answer may not accurate.

Part 2

Data

This part is done with `tidyverse` (Wickham et al.,2019)

Simulation process

I assume that there are 10000 students in the U of T and I only choose 20 from each stratified goup in total 1000 observations as my stratified sampling. Since I use the stratified sampling, I first simulate the population by doing the below steps: 1. First, I use the set.seed method to simulate as the same result after different times of simulation. Then, I randomly choose 10000 students and give them the id from 1 to 10000 and store the data as variable `number_students`. Since I use distance to classify different groups, I randomly simulate 200 distances from 1 km to 50 km by using method `rep` which means repeatedly choose and name the variable as `distance`. Then, I create a variable to store all the possible modes of transportation such as `car`, `driving` and so on and then I simply choose a size of 10000 population with replacement to simulate and store the value as `actual_mode`. Similarly, the variable `favor_mode` does the similar way as `actual_mode` by simulating 10000 times with replacement under different modes of transportation. As for variable `commute_time`, I simply choose 10000 times randomly with replacement under the time range 10 to 100 minutes including 500 NA values. To simulate if students like in person classes or not, I create a variable called "in_person" and base on the response `Yes` or `No` to simulate a size of 10000 population with replacement by using `sample`. Then to simulate how often a student go for school once a week, a variable named `day_to_school` store the simulation of 10000 population under the mean is 3 and standard deviation is 0.5 by using `rnorm`(under normal distribution). Then to simulate the times student late for school, I also create 10000 population with lambda is 3 and use method `rpois` (Poison distribution) and store the value into variable `late_time_school`. Since in the survey, I make a short answer for suggestion question, then here I create a population by simulate if or not students write response. Thus, `suggestion` is to store the response after simulate 10000 population with response `Yes` or `No` with replacement. As for the sex, in my survey, I design as short answer, but now I just simulate 10000 times with replacement to get a population with three choices `Male`, `Female` and `Other` and store the result of simulation into `sex`. Also different from the survey, instead of short answer, I create a list of age range from 16 to 24 and named the list as variable `age_students`. Then still do the simple random sampling to choose 10000 times from the `age_students` with replacement and store the results as variable `age`. Finally, after getting all the information by simulation, I create a variable called `summary_info` to store all the results into a table by using method `tibble`. Then, I can perform stratified method to classify the population as different groups according to the distance(`group_by(distance)`) from home to school and them I select 20 samples(`sample_n(20)`) from each group and make each select results in total 1000 observations as a sample. That is the whole process for my data simulation.

Clean data

I also simulated some NA value which I store as `simulate_na`. And I assume there are 500 NA in variable `commute_time`. Just simulate in a real situation, some students may not ask for this question that will happen NA cases. Then in order to clean the NA value, after I get the `summary_info` I use `filter` to filter out the rows with NA and then do the stratified sampling.

The first important variable is `actual_mode`. This variable stores the mode of transportation that students use to go to school. This data is simulated by R and there are 10000 observations of students who actual take which mode of transportation to go to school. It is also a categorical variable. The second important variable is `favor_mode`. This variable stores the mode of transportation that students want to use to go to school but they may not use due to some special reasons. This data is simulated by R and there are 10000 observations of students who want to take which mode of transportation to go to school. It is also a categorical variable. The third important variable is `day_to_school`. This variable stores how many days those students go to school once a week. This data is simulated by R and there are 10000 observations of students who go to school how many times once a week. It is a numeric variable typically range from 0 to 7 but may happens some outliers. The forth important variable is `late_school_time`. This variable stores how many times that students go to school late once a week. This data is simulated by R and there are 10000 observations of

students who late for school how many times once a week. It is a numeric variable typically range from 0 to 7 but may happens some outliers.

Table 1: actual transportation

least_actual	max_actual
bicycle	walking

Table 2: favor transportation

least_favor	max_favor
bicycle	walking

Summary

This part is done with `kableExtra` (Hao Zhu (2021))

As for the `summary_actual` and `summary_favor`. I both get the max and min to see which modes are the most or least popular for students to choose to go to school. It is clear that most students prefer walking to school and they actually do walking to school. But bicycle is not very popular for students. Thus, U of T should pay more attention to construct more walking path as soon as possible to fit most students choice. On the other hand, U of T can also construct more bicycle parking area that may let more students choose to ride to school.

For the `summary_go_school`. I choose to calculate the median and mean since I think they are more representative than other data like min or max and so on. The mean for students to go to school is nearly 3 days(3.0178278) once a week and the median is quite similar which is also nearly 3 days(3.0107282). That means students in U of T almost have to go to school for classes 3 times a week. Because of this kind of frequent trips to school, school should try to improve students' commuting problems as much as possible to save students' commuting time. Suggestions can just follow my previous analysis as `summary_actual` and `summary_favor`, after solving these problems, students may be more satisfy.

I also analysis `summary_late_school` by taking the mean and median of this data. Standard deviation is also useful in this case. From the value of mean and median, I notice that it is quite similar as `summary_late_school`. The mean is 3.009 and median(3) is 3 for this case which means many students late for school quite often. Since the mean and median are both around 3 which means that I can use this data to do the Hypothesis test since it meets the assumption of normality. Combine `summary_late_school` and `summary_go_school`, it shows that many students late for school nearly every single day. But, since the standard deviation is large which is 1.7087536 that means the variation is large between different students. Some students may late for school much less than others, vice versa. In conclusion, based on the previous summaries, U of T does need to come up with some new decisions to decrease the lateness rate for students. Especially for walking path construction.

Table 3: Day to school

go_school_mean	go_school_median
3.017828	3.010728

Table 4: Late to school

late_school_mean	late_school_median	late_school_sd
3.009	3	1.708754

Figure1: How many times students go to school once a week

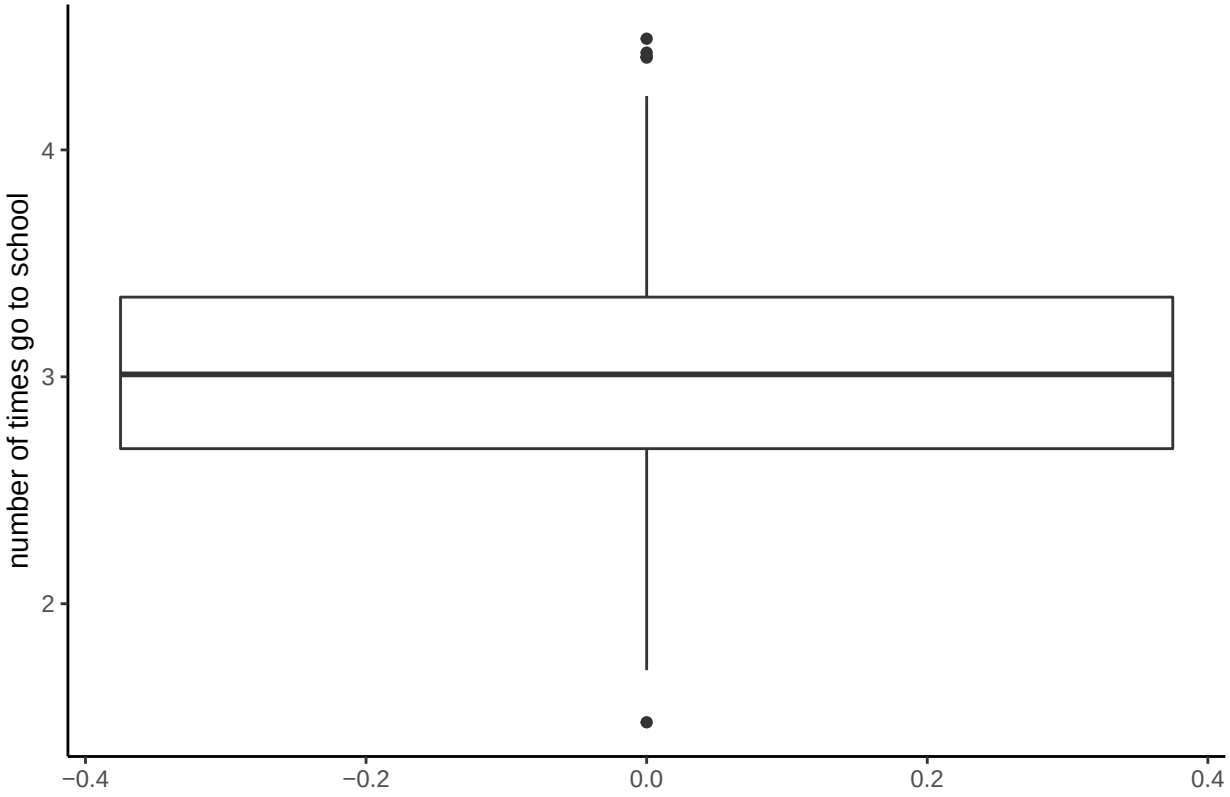


Figure2: How many times students go to school late once a week

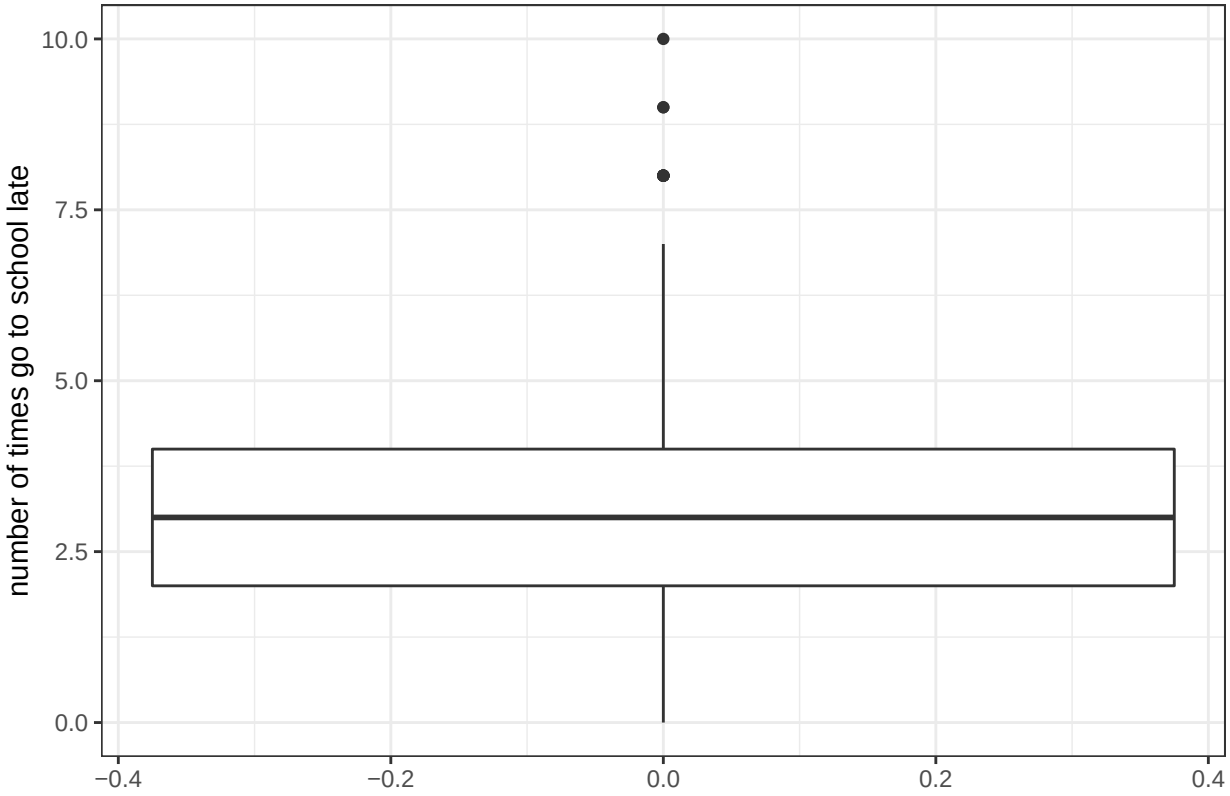
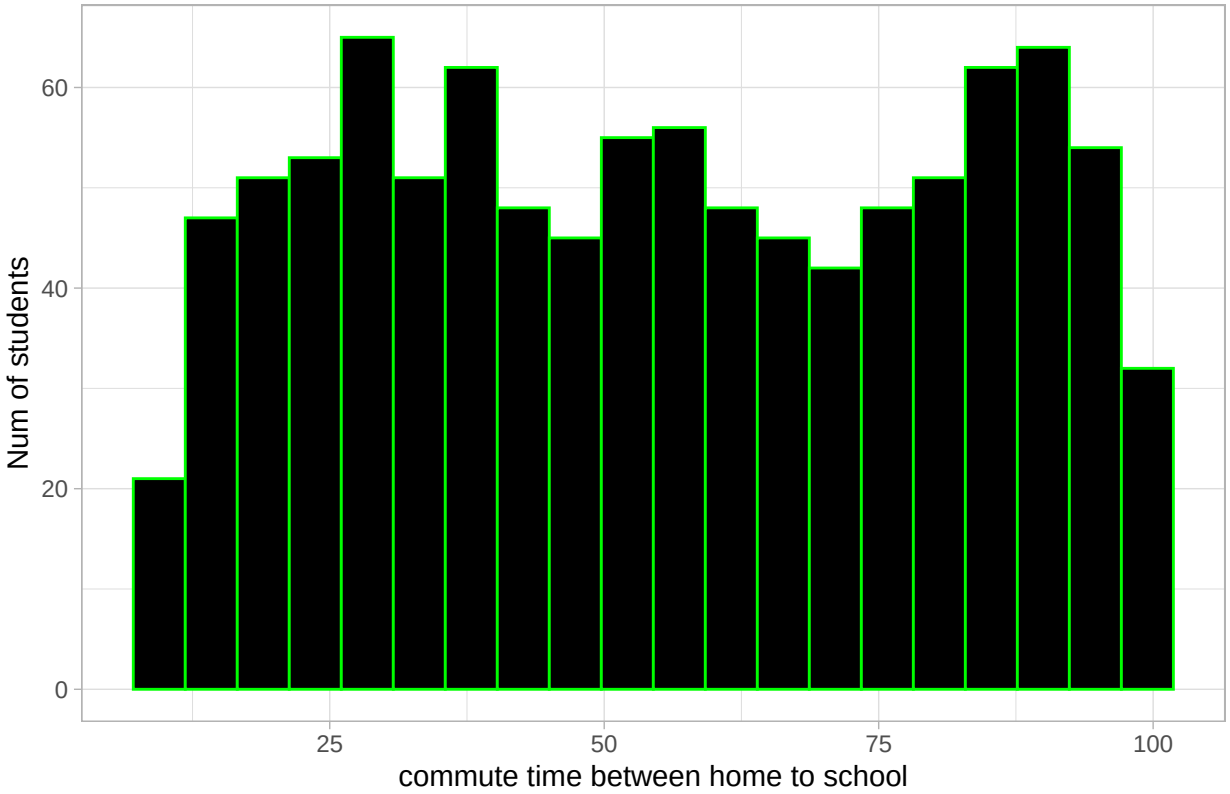


Figure3: Histogram of commute time (in minutes)per students



The first boxplot(Figure 1) shows how many times students go to school once a week. The median is nearly at 3 and the IQR nearly ranges from 2.5 to 3.5. There are also many outliers from the two sides of graph which the bigger outliers range from nearly 4.8 to 5.2 and the smaller outlier is about 0.8.

The second boxplot(Figure 2) shows how many times students late for school once a week. The median is nearly at 3 and the IQR nearly ranges from 2 to 4. There are also three outliers which are much bigger than other data. Range nearly from about 8 to 10. Since the actual max days late for school is 7, thus a few data shows in the graph are bad simulation date.

The histogram(Figure 3) shows how much time students need for commuting. This histogram is nearly symmetric with mode at about 55 minutes. The outlier is clear to see that both on the most right and left side of the histogram. Also I can assume that the commute time for U of T students is average, with some students being basically evenly distributed across time periods.

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

Methods

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{(n)}}}$$

I am interested in the parameter `late_time_school` which is a numerical variable. It means how many days U of T students will late for school once a week. First, I assume that the confidence interval follows the normal distribution. Since the for the sample mean I simulated for `late_time_school` is 3.009 which is the estimation of U of T students who late for school 3 days in a week. In order to get more specific information, I need to make a bootstrap to construct a 99% confidence interval to present a range that has 99% confidence to say the true mean time for U of T students late for school is in this range. Then I will simulate the distribution of the sample mean that have middle 99% range of the possible sample means in

Table 5: Summary

variable	summary_val
CI_lower	2.8649700
CI_upper	3.1315050
p-value	0.8677523

the results part. First, creating a variable called `late` to store the simulated mean value of late times by using a while loop to loop for 500 times. Then, make a table to show the simulated mean of late time and use the method `quantile` to get the 99% confidence interval.

Hypothesis test

I will assume the mean of `late_time_school` be $H_0 = 3.009$ which is null hypothesis and H_a is the mean of `late_time_school` is not 3.009. I first get the total observation of `late_time_school` and store as `row` and then by using the formula I present above to calculate the t value which is `(late_summary$late_mean - H0) / (late_summary$late_sd /sqrt(nrow(strata))` in code. Since I get the t value and I want to get the p value to test if I have evidence to against H_0. Since I want to test if H_0 = `summary_late_school$late_school_mean` I recall the formula I learnt in STA238 which follow the code '2 * (1 - pt(t, degree freedom)). For here, the degree freedom is 1 and the total observation is row, then the degree freedom is row - 1. After I get the p value I make a summary to combine the pvalue and confidence interval and then use "knitr::kable" to make a table for the summary

Results

The result I calculated by 99% bootstrap confidence interval is (2.86497, 3.131505) this result means that we have 99% confidence the actual late for school time for U of T students is in this interval. This result is quite reasonable since I have calculated in the data section that students mean late for school times is 'r 3.009 which falls in this range. Since the average late days for class is quite high once a week. This result also mentions that universities should build more convenient facilities for students to try the best reduce the lateness rate for U of T students. Also, the confidence interval is relatively large due to the small sample, if the sample is large enough, the confidence interval will be smaller.

Since I assume that $H_0 = 3.009$ and $H_a \neq 3.009$ which both H_0 and H_a I have mentioned in the method section. The p-value I calculated as 0.8677523. Since the p-value is much larger than practically significant which is 0.1, that means we have no evidence to against the null Hypothesis H_0. Thus, by doing the hypothesis test, this result also supports that the mean late times for U of students is about 3 times per week. Thus, U of T should offer more convenient facilities for students to commute. Like building more walking path, constructing more parking places, offering more school buses and ask for more TTC paths that can go to U of T directly.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/.
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: May 5, 2021)
4. Hao Zhu (2021) *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra https://github.com/haozhu233/kableExtra.

5.JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2022) *rmarkdown: Dynamic Documents for R. R package version 2.16.* https://rmarkdown.rstudio.com.

6. Unkown. University of Toronto.(2022) https://future.utoronto.ca/university-life/transportation/

7. Wickham et al. (2019) *"Welcome to the tidyverse."* URL: https://doi.org/10.21105/joss.01686

8. Yihui Xie (2022) *knitr: A General-Purpose Package for Dynamic Report Generation in R.*

9.https://rmarkdown.rstudio.com/lesson-7.html

Appendix

Here is a glimpse of the data set simulated:

```
## Rows: 1,000
## Columns: 11
## $ number_students  <int> 3351, 2101, 2501, 4351, 1401, 4901, 3001, 7251, 7451,~
## $ distance         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ actual_mode      <chr> "uber", "walking", "school bus", "driving", "school b~
## $ commute_time     <int> 24, 69, 38, 57, 81, 16, 97, 36, 39, 76, 30, 33, 40, 6~
## $ in_person        <chr> "No", "Yes", "No", "No", "No", "No", "Yes", "Yes", "Y~
## $ day_to_school    <dbl> 3.576817, 2.149842, 3.523391, 2.956233, 3.035240, 2.8~
## $ late_time_school <int> 2, 1, 3, 1, 6, 6, 1, 2, 2, 2, 4, 7, 2, 5, 6, 6, 5, 0,~
## $ suggestion       <chr> "No", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes",~
## $ sex              <chr> "Female", "Other", "Male", "Female", "Male", "Male", ~
## $ age              <dbl> 23, 21, 16, 24, 17, 24, 22, 24, 19, 16, 20, 16, 19, 1~
## $ favor_mode       <chr> "car", "school bus", "TTC", "skating", "school bus", ~
```