1 Introduction

1.1 Background

Nowadays, with the explosion of population in mega cities, traffic in cities centers is increasingly worse. To solve this phenomenon, people start explore new ways of transportation to mitigate the pressure or traffic. Among these attempts, bike sharing is an impressive one. The idea of bike sharing is eco-friendly, feasible and convenient. Bike sharing has proved to be successful in many big cities around the world. As the third-most populous in the United States, Chicago has build massive public transportation system, among which sharing bike is a irreplaceable part.

In the program of the Chicago Department of Transportation (CDOT), Divvy is the bike share system, providing residents and visitors with a convenient, fun and affordable transportation option for getting around and exploring Chicago. Originally, Divvy provide classic and docked bike, which has to be parked in thousands of stations. Recently, electric bicycles like scooter are introduced, which can be parked anywhere within certain region, which further promote the convenience of this bike sharing system.

1.2 Objective

As the bike sharing system across Chicago downtown, Divvy's mission is to keep the city's surface transportation networks and public ways safe for users, environmentally sustainable and in a state of good repair and attractive, so that its diverse residents, businesses and guests can all enjoy a variety of quality transportation options, regardless of ability or destination. So, we would like to examine the detailed effect of this system, by our first parameter of interest. By estimating **the average time a ride take**, we can see whether majority users truly use the bike to commute, supplanting other transportation methods or just use it for connection between subway and bus, etc.

Beyond that, we also care about the popularity of the new electric bikes. We want to explore whether electric bike attract more users than other traditional one. We assess this by our second parameter of interest, which is **the proportion of rides using electric bicycles**.

2 Data Collection and Summaries

2.1 About the Dataset

For this project, we approximately treat the data as the whole population as our data contains more than 200000 observations.

Our dataset records comprehensive information about the bike rides taken by users of Divvy, a bike-share company based in Chicago. The data has been made available by Motivate International Inc. with license here. The raw dataset contains 12 sub-datasets, which contain the information in each month during the period *May 2022 - April 2023.*

We chose the data collected in March 2023 as our subject. There exists some NA's and empty values in the data. Therefore we first clean the data and drop all inappropriate rows. In addition, as one of our interest parameter is the mean ride duration of all users, we added a new column called 'duration' by subtracting the end time by the start time in the table and then converted it to numeric value in minutes.

After all the wrangling operations, we found that some rides only take less than one minute, which is abnormal and may indicate the bicycle is broken. These extremely short ride times can have a profound bias impact on the mean ride duration which is one of our interested parameter. Hence we delete all rides with duration time less than one minutes. We also add a column of displacement between the starting and ending stations, serving as the auxiliary variable.

Detail variable names and interpretations are listed in the table below:

Variable Name	Interpretation
ride_id	ID recording a ride
rideable_type	Types of bike rides available
started_at	Time the ride started
ended_at	Time the ride ended
start_station_name	Where the ride started
start_station_id	ID of the initial stations
end_station_name	Where the ride ended
end_station_id	ID of the ending stations
start_lat	Coordinates, starting latitude
start_lng	Coordinates, starting longitude
end_lat	Coordinates, ending latitude
end_lng	Coordinates, ending longitude
member_casual	Whether a annual member or not
**duration*	Duration of a ride (in minutes)
**displacement*	Displacement between the starting and ending stations

* : variable that we created based on other variables in the raw dataset.

The target population: all rides of Divvy in Chicago in March 2023

Parameter of interest:
1. The mean ride duration all rides of Divvy in Chicago in March 2023
2. The proportion of electric bike in all rides

2.2 Sampling Method and Procedure

In this part, we will use two sampling methods: **Simple Random sampling** and **Stratified Sampling**. In the following section, we will describe our auxiliary variable, sample size, strata size and the detailed sampling procedure.

2.2.1 Auxiliary Variable

We decided to use both vanilla and ratio estimate to estimate our parameter of interest. So an auxiliary variable need to be decide. However, only the starting time and finishing time are continuous, which have been used already to compute the duration of every ride. So we created new continuous variable – "displacement", which is the direct distance between the starting position and the end position. The reason of this auxiliary variable is that the larger the distance, the longer a ride takes usually.

2.2.2 Sample Size

To get a reasonable sample size, we use the 95% confidence interval

$$n_0 = \frac{(z_{\alpha/2})^2 * p_{guess} * (1 - p_{guess})}{\sigma^2} = \frac{1.96^2 * 0.5 * (1 - 0.5)}{0.05^2} = 384.16$$

$$n = \frac{n_0}{1+n_0/N} = \frac{384.16}{1+384.16/194432} = 383.4025 \rightarrow n = 384$$

2.2.3 Strata Size in Stratified Sampling

In the data, users of each ride are divided into member and non-member. We believe that the variance between the ride time of annual member and that of casual user can be very large. So we draw two strata here, one sampled from rides by member users and another sampled from rides by casual users. In terms of the sample size, we used proportional allocation:

$$\frac{n_m}{N_m} = \frac{n_c}{N_c}$$

Where:
n_m : the number of member users in strata m
N_m : the number of member users in all rides
n_c : the number of casual users in strata c
N_c : the number of casual users in all rides

With the constraint $n_m + n_c = n = 384,\ N_m = 148773,\ N_c = 45730,\ n_m = 294,\ n_c = 90$

3 Data Analysis

In this section, for each of our two parameters of interest, we use two sampling method, SRS and stratified sampling. In the SRS, we use both vanilla and ratio estimator.

3.1 Simple Random Sampling for mean ride duration

3.1.1 Using Vanilla Estimate:

Estimate:
$\bar{y}_s = 11.04488$

Standard Error:
$SE(vanilla) = \sqrt{(1 - \frac{n}{N}) \times \frac{s_s^2}{n}} = \sqrt{(1 - \frac{384}{194503}) \times \frac{148.1839}{384}} = 0.6205915$

95% Confidence Interval:
$\bar{y}_s \pm 1.96 \times SE(vanilla) = [9.8285192,\ 12.2612378]$

3.1.2 Using Ratio Estimate:

Estimate:
$\frac{\bar{y}_s}{\bar{x}_s} \times \bar{x}_p = 11.1010304$

Standard Error:
$s_e^2 = \frac{1}{n-1}\sum_{i \in S} e_i^2 = \frac{1}{n-1}\sum_{i \in S}(y_i - \frac{\bar{y}_s}{\bar{x}_s} \times x_i)^2 = 127.7312$

$SE(ratio) = \sqrt{(1 - \frac{n}{N}) \times \frac{s_e^2}{n}} = \sqrt{(1 - \frac{384}{194503}) \times \frac{127.7312}{384}} = 0.5761741$

95% Confidence Interval:
$\frac{\bar{y}_s}{\bar{x}_s} \times \bar{x}_p \pm 1.96 \times SE(ratio) = [9.9717291,\ 12.2303317]$

3.1.3 Interpretation:

Our assumption is to treat our data as the population, so we assume knowing the population size and the population mean of our auxiliary variable.

For the vanilla estimator, the estimate of the true mean of ride duration is 11.04488 minutes, with standard error 0.6205915. We are 95% confident that in 19 out of 20 resamplings, the interval [9.8285192, 12.2612378] will capture the true mean of ride duration.

For the ratio estimator, the estimate of the true mean of ride duration is 11.1010304 minutes, with standard error 0.5761741. We are 95% confident that in 19 out of 20 resamplings, the interval [9.9717291, 12.2303317] will capture the true mean of ride duration.

Compare these two estimates, ratio estimate has smaller standard error, meaning that ratio estimator works better than vanilla estimator. This implies that the correlation between our response mean of ride duration and our auxiliary variable displacement is strong, and our choice of auxiliary variable is reasonable.

3.2 Stratified Sampling for mean ride duration

3.2.1 Estimate:

$\hat{y}_{str} = \sum_{h=1}^{2}(\frac{N_h}{N})\hat{y}_{Sh} = \frac{148773}{194503} \times 10.08362 + \frac{45730}{194503} \times 16.62667 = 11.62197$

3.2.2 Standard Error:

$SE(\hat{y}_{Sm}) = \sqrt{(1 - \frac{n_m}{N_m})\frac{s_{Sm}^2}{n_m}} = \sqrt{(1 - \frac{294}{148773}) \times \frac{8.804082^2}{294}} = 0.5129569$

$SE(\hat{y}_{Sc}) = \sqrt{(1 - \frac{n_c}{N_c})\frac{s_{Sc}^2}{n_c}} = \sqrt{(1 - \frac{90}{45730}) \times \frac{21.88502^2}{90}} = 2.304613$
$SE(\hat{y}_{str}) = \sqrt{\sum_{i \in \{m,c\}}(\frac{N_i}{N})^2 SE^2[\hat{y}_{Si}]} = \sqrt{(\frac{148773}{194503})^2 \times 0.5129569^2(\frac{45730}{194503}) \times 2.304613^2}$
$SE(\hat{y}_{str}) = 0.6689807$

3.2.3 95% Confidence Interval:

$\hat{y}_{str} \pm 1.96 \times SE(\hat{y}_{str}) = [10.95299,\ 12.29095]$

3.2.4 Interpretaion:

The stratified estimate of the true mean of ride duration is 11.62197, with standard error 0.6689807. We are 95% confident that in 19 out of 20 resamplings, the interval [10.95299, 12.29095] will capture the true proportion of electric bike in all rides.

Compared with SRS the standard error of stratified estimate is even larger, which means SRS works better here. This implies the variation between two strata member and non-member is small. The reason could be that for no matter member or casual user, everyone use the sharing bike only in short commute, maybe member use the bike more often, but the time cost by each time is similar to that of non-member users.

3.3 Simple Random Sampling for proportion of electric bike

3.3.1 Using Vanilla Estimate:

Estimate:
$\hat{p} = \frac{n_e}{n} = \frac{172}{384} = 0.4479167$

Standard Error:

$$SE(\hat{p}) = \sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{(1-\frac{384}{194503})\frac{0.4479167\times(1-0.4479167)}{384}} = 0.02535165$$

95% Confidence Interval:
$\hat{p} \pm 1.96 \times SE(\hat{p}) = [0.244, 0.334]$

3.3.2 Using Ratio Estimate:

Estimate:
$\hat{p}_p = \frac{\hat{p}_s}{\bar{x}_s}\bar{x}_p = \frac{0.4479167}{1892.308} * 1863.293 = 0.4501939$

Standard Error:
$s_e^2 = \frac{\sum_{i\in S} e_i^2}{n-1} = \frac{\sum_{i\in S}(z_i - \frac{\hat{p}_s}{\bar{x}_s}x_i)^2}{n-1}$

$$z_i = \begin{cases} 1, & \text{if ith ride using Electric bike} \\ 0, & \text{otherwise} \end{cases}$$

$$SE(\hat{p}_p) = \sqrt{(1-\frac{n}{N})\frac{s_e^2}{n}} = \sqrt{(1-\frac{384}{194503})\times\frac{0.3799694}{384}} = 0.03142531$$

95% Confidence Interval:
$\hat{p}_p \pm 1.96 \times SE(\hat{p}_p) = [0.389, 0.512]$

3.3.3 Interpretation:

Our assumption is to treat our data as the population, so we assume knowing the population size and the population mean of our auxiliary variable.

For the vanilla estimator, the estimate of the true proportion of electric bike in all rides is 0.4479167, with standard error 0.02535165. We are 95% confident that in 19 out of 20 resamplings, the interval [0.244, 0.334] will capture the true proportion of electric bike in all rides.

For the ratio estimator, the estimate of the true proportion of electric bike in all rides is 0.4501939, with standard error 0.03142531. We are 95% confident that in 19 out of 20 resamplings, the interval [0.389, 0.512] will capture the true proportion of electric bike in all rides.

Here the two estimates are very close. The vanilla method works better with smaller standard error. This means that the correlation between our response proportion of electric bike in all rides and our auxiliary variable displacement is weak. This result is expected since the distribution of different types of bikes are mainly is random: users may not be able to choose electric bikes or others, just use whichever type of bicycles in front of them instead.

3.4 Stratified Sampling for proportion of electric bike

3.4.1 Estimate:

$\hat{y}_{str} = \sum_{i \in \{m,c\}} (\frac{N_i}{N}) \hat{p}_i = 0.4167255$

3.4.2 Standard Error:

$SE(\hat{p}_m) = \sqrt{(1 - \frac{n_m}{N_m}) \frac{\hat{p}_m(1-\hat{p}_m)}{n_m}}$
$SE(\hat{p}_c) = \sqrt{(1 - \frac{n_c}{N_c}) \frac{\hat{p}_c(1-\hat{p}_c)}{n_c}}$
$SE(\hat{y}_{str}) = \sqrt{\sum_{i \in \{m,c\}} (\frac{N_i}{N})^2 SE^2(\hat{p}_i)} = 0.02507541$

3.4.3 95% Confidence Interval

$\hat{y}_{str} \pm 1.96 \times SE(\hat{y}_{str}) = [0.392, 0.442]$

3.4.4 Interpretation:

The stratified estimate of the true proportion of electric bike in all rides is 0.4167255, with standard error 0.02507541. We are 95% confident that in 19 out of 20 resamplings, the interval [0.392, 0.442] will capture the true proportion of electric bike in all rides.

Compared with SRS, the standard error of stratified estimate is smaller than ratio estimate, but very close to vanilla estimate. This means that the with-in strata variance is not really large as we expected. Hence the stratified sampling failed to perform better than simple random sampling. The reason could be that the variation between member and non-member on the choice of bike type is relatively small: all users just take bike randomly.

4 Conclusion and Discussion

4.1 Conclusion

After conducting two sampling method and three estimator, we get preliminary estimate of our parameters of interest. In conclusion, the true mean of the duration per ride is 8.33544 minutes, considered accurate within 66.89807% points, 19 out of 20 times. The true proportion of electric bike in all rides is 41.67255%, considered accurate within 2.507541% points, 19 out of 20 times.

Back to our objectives. Clearly, most users only use the sharing bike for short commute in the city (less than 10 minutes) and the bike sharing system currently cannot replace other public transportation in terms of long distance transit. Electric bike is indeed more popular than other two types of traditional bicycles, with the estimated proportion of 41.67

4.2 Discussion

4.2.1 Limitation

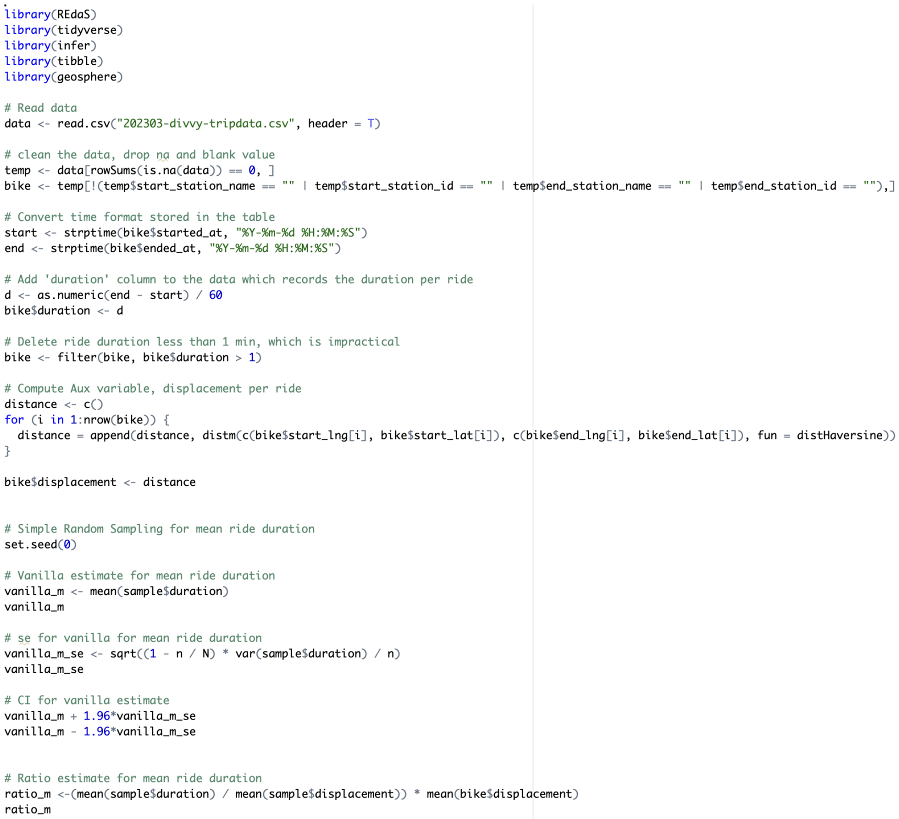
There exists a variety of limitation in this study. Firstly, to conduct sampling and estimating parameters of interest, we have made several assumption, and hence once without the condition, the conclusion may not be applicable. In addition, in terms of the dataset, the available information of riders and ride is not enough for us to draw a persuasive conclusion, and we also have no idea about the missing and empty value in the raw dataset. The lack of information may lead to biased result.

4.2.2 Further Application

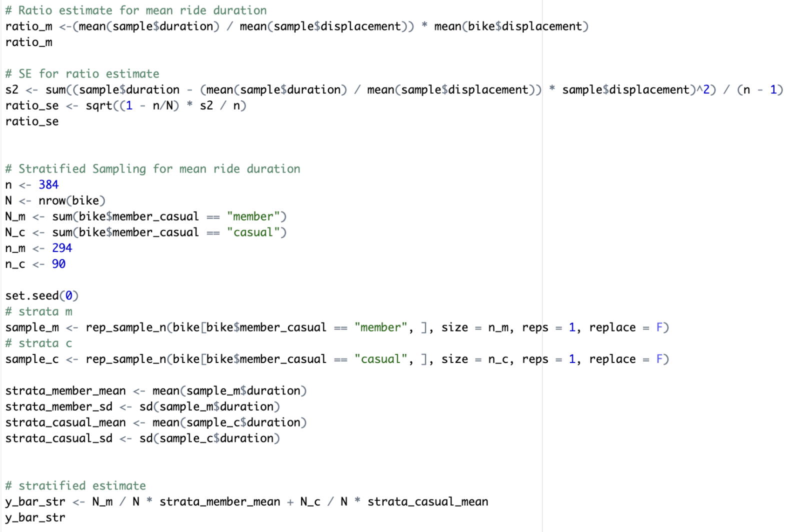
Our current study only focus on the estimation and interpretation of two parameters of interest. In the future, it can be enriched by other statistical analysis, such as hypothesis testing or model prediction. In addition, the objective can be extended to other scope about this bike sharing system since we have only use a few columns of information in the raw data. Eventually, the raw dataset actually contains 12 sub-datasets, which record the information in each month during the period May 2022 - April 2023. Therefore, time series can be applied to the whole large dataset to explore certain trend.

Appendix

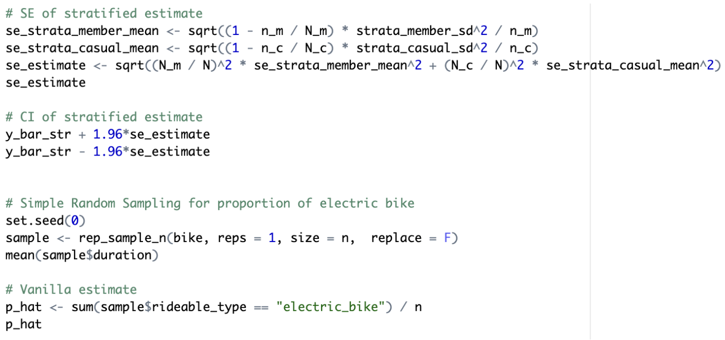
R code



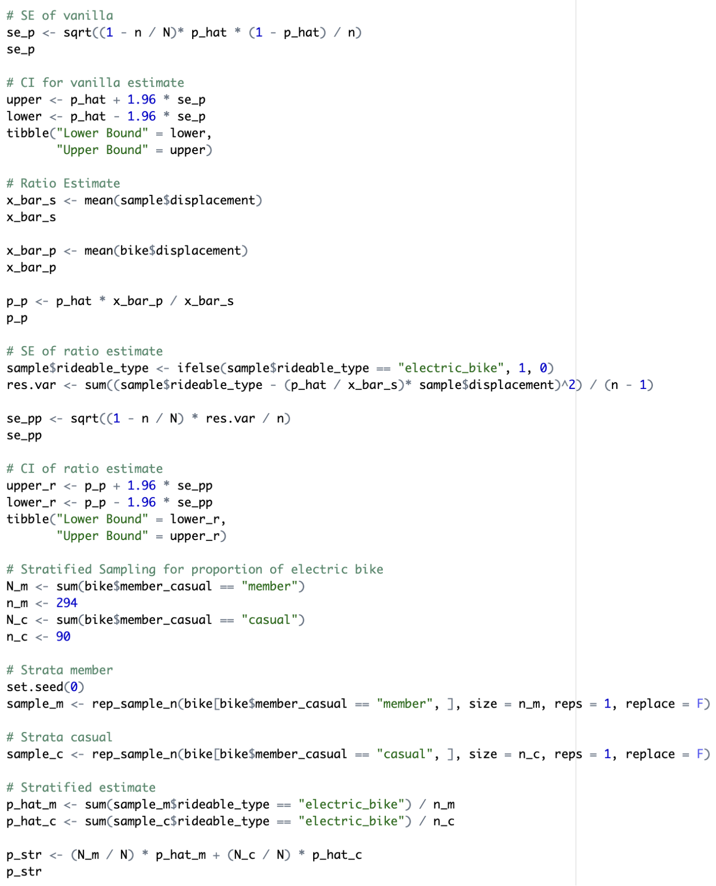
```
library(REdaS)
library(tidyverse)
library(infer)
library(tibble)
library(geosphere)

# Read data
data <- read.csv("202303-divvy-tripdata.csv", header = T)

# clean the data, drop na and blank value
temp <- data[rowSums(is.na(data)) == 0, ]
bike <- temp[!(temp$start_station_name == "" | temp$start_station_id == "" | temp$end_station_name == "" | temp$end_station_id == ""),]

# Convert time format stored in the table
start <- strptime(bike$started_at, "%Y-%m-%d %H:%M:%S")
end <- strptime(bike$ended_at, "%Y-%m-%d %H:%M:%S")

# Add 'duration' column to the data which records the duration per ride
d <- as.numeric(end - start) / 60
bike$duration <- d

# Delete ride duration less than 1 min, which is impractical
bike <- filter(bike, bike$duration > 1)

# Compute Aux variable, displacement per ride
distance <- c()
for (i in 1:nrow(bike)) {
  distance = append(distance, distm(c(bike$start_lng[i], bike$start_lat[i]), c(bike$end_lng[i], bike$end_lat[i]), fun = distHaversine))
}

bike$displacement <- distance


# Simple Random Sampling for mean ride duration
set.seed(0)

# Vanilla estimate for mean ride duration
vanilla_m <- mean(sample$duration)
vanilla_m

# se for vanilla for mean ride duration
vanilla_m_se <- sqrt((1 - n / N) * var(sample$duration) / n)
vanilla_m_se

# CI for vanilla estimate
vanilla_m + 1.96*vanilla_m_se
vanilla_m - 1.96*vanilla_m_se


# Ratio estimate for mean ride duration
ratio_m <-(mean(sample$duration) / mean(sample$displacement)) * mean(bike$displacement)
ratio_m
```



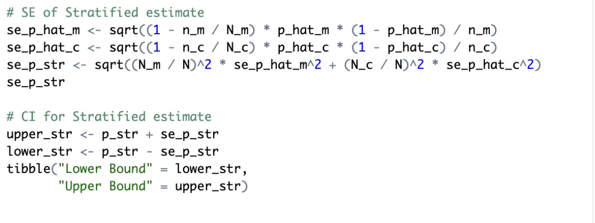
```
# Ratio estimate for mean ride duration
ratio_m <-(mean(sample$duration) / mean(sample$displacement)) * mean(bike$displacement)
ratio_m

# SE for ratio estimate
s2 <- sum((sample$duration - (mean(sample$duration) / mean(sample$displacement)) * sample$displacement)^2) / (n - 1)
ratio_se <- sqrt((1 - n/N) * s2 / n)
ratio_se


# Stratified Sampling for mean ride duration
n <- 384
N <- nrow(bike)
N_m <- sum(bike$member_casual == "member")
N_c <- sum(bike$member_casual == "casual")
n_m <- 294
n_c <- 90

set.seed(0)
# strata m
sample_m <- rep_sample_n(bike[bike$member_casual == "member", ], size = n_m, reps = 1, replace = F)
# strata c
sample_c <- rep_sample_n(bike[bike$member_casual == "casual", ], size = n_c, reps = 1, replace = F)

strata_member_mean <- mean(sample_m$duration)
strata_member_sd <- sd(sample_m$duration)
strata_casual_mean <- mean(sample_c$duration)
strata_casual_sd <- sd(sample_c$duration)


# stratified estimate
y_bar_str <- N_m / N * strata_member_mean + N_c / N * strata_casual_mean
y_bar_str
```



```r
# SE of stratified estimate
se_strata_member_mean <- sqrt((1 - n_m / N_m) * strata_member_sd^2 / n_m)
se_strata_casual_mean <- sqrt((1 - n_c / N_c) * strata_casual_sd^2 / n_c)
se_estimate <- sqrt((N_m / N)^2 * se_strata_member_mean^2 + (N_c / N)^2 * se_strata_casual_mean^2)
se_estimate

# CI of stratified estimate
y_bar_str + 1.96*se_estimate
y_bar_str - 1.96*se_estimate


# Simple Random Sampling for proportion of electric bike
set.seed(0)
sample <- rep_sample_n(bike, reps = 1, size = n,  replace = F)
mean(sample$duration)

# Vanilla estimate
p_hat <- sum(sample$rideable_type == "electric_bike") / n
p_hat
```



```r
# SE of vanilla
se_p <- sqrt((1 - n / N)* p_hat * (1 - p_hat) / n)
se_p

# CI for vanilla estimate
upper <- p_hat + 1.96 * se_p
lower <- p_hat - 1.96 * se_p
tibble("Lower Bound" = lower,
       "Upper Bound" = upper)

# Ratio Estimate
x_bar_s <- mean(sample$displacement)
x_bar_s

x_bar_p <- mean(bike$displacement)
x_bar_p

p_p <- p_hat * x_bar_p / x_bar_s
p_p

# SE of ratio estimate
sample$rideable_type <- ifelse(sample$rideable_type == "electric_bike", 1, 0)
res.var <- sum((sample$rideable_type - (p_hat / x_bar_s)* sample$displacement)^2) / (n - 1)

se_pp <- sqrt((1 - n / N) * res.var / n)
se_pp

# CI of ratio estimate
upper_r <- p_p + 1.96 * se_pp
lower_r <- p_p - 1.96 * se_pp
tibble("Lower Bound" = lower_r,
       "Upper Bound" = upper_r)

# Stratified Sampling for proportion of electric bike
N_m <- sum(bike$member_casual == "member")
n_m <- 294
N_c <- sum(bike$member_casual == "casual")
n_c <- 90

# Strata member
set.seed(0)
sample_m <- rep_sample_n(bike[bike$member_casual == "member", ], size = n_m, reps = 1, replace = F)

# Strata casual
sample_c <- rep_sample_n(bike[bike$member_casual == "casual", ], size = n_c, reps = 1, replace = F)

# Stratified estimate
p_hat_m <- sum(sample_m$rideable_type == "electric_bike") / n_m
p_hat_c <- sum(sample_c$rideable_type == "electric_bike") / n_c

p_str <- (N_m / N) * p_hat_m + (N_c / N) * p_hat_c
p_str
```



```r
# SE of Stratified estimate
se_p_hat_m <- sqrt((1 - n_m / N_m) * p_hat_m * (1 - p_hat_m) / n_m)
se_p_hat_c <- sqrt((1 - n_c / N_c) * p_hat_c * (1 - p_hat_c) / n_c)
se_p_str <- sqrt((N_m / N)^2 * se_p_hat_m^2 + (N_c / N)^2 * se_p_hat_c^2)
se_p_str

# CI for Stratified estimate
upper_str <- p_str + se_p_str
lower_str <- p_str - se_p_str
tibble("Lower Bound" = lower_str,
       "Upper Bound" = upper_str)
```