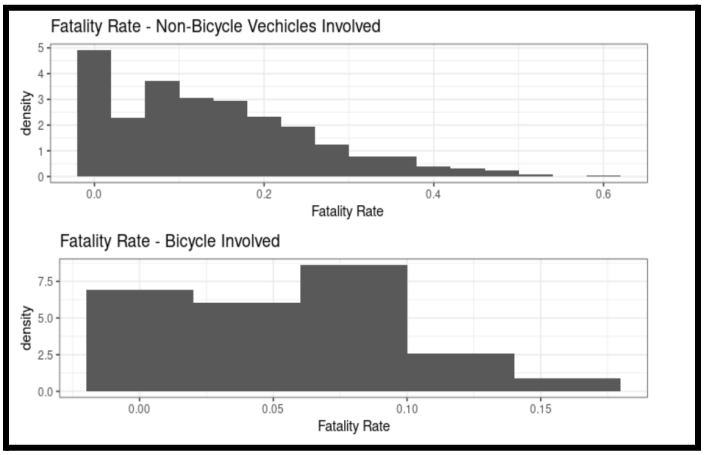
MVC Data Analysis

Citizens of big metropolicies all around the globe are switching from convenient and conventional cars to bicycles as their main form of transportation. For example, in the Netherlands, bicycles are used for around 27% of all trips [2]. There are many reasons for it, bicycles are cheaper and better for the environment, but are they as safe as other vehicles? This research will be focused on applying the statistical methodologies to the collision's data in order to answer the following research question: *Does the true proportion of fatal injuries in the collisions involving bicycles differ from the true proportion of fatal injuries in the collisions not involving bicycles?*. In order to perform the statistical analysis, we will be working with "Motor Vehicle Collisions involving Killed or Seriously Injured Persons" dataset [1]. MVC dataset was collected and summarized by Toronto Police Services. It provides a comprehensive overview of all traffic collision events where a person was either Killed or Seriously Injured (KSI) from 2006 – 2020 in the city of Toronto.

Data: MVC dataset has mainly qualitative variables like the location and time of the collision (ex. DATE, STREET1), classification of the incident and injuries (ex. ACCLASS), information about what vehicles were involved (ex.VEHTYPE) and etc. Two variables will be used to compute the rate of fatal injuries based on the vehicle type. The first variable that will be used is the information about the injury: (ACCLASS) which can be "Fatal" or "Not Fatal", and will be used to separate the data into 2 groups. The second variable: (VEHTYPE) demonstrates the type of vehicles involved (ex. "Bicycle" - motorcycle was involved in the collision) and will be used to compute the number of collisions that involve bicycles.

Additional data-cleaning will be performed. At first and all data points with N/A values in the aforementioned variables will be removed. In addition, the quantitative binary data of ACCLASS variables will be changed to 1 or 0 ("Fatal" = 1 or "Not Fatal" =0), to simplify the computation of statistics. For EDA analysis data will be divided into 2 groups: data1 - all collisions and data2 - only motorcycle-involved collisions based on the VEHTYPE variable.

EDA: Sample size is large enough to work with, since there are 13325 data-points of non-bicycle collisions and 722 data-points of bicycle collisions. Based on the simulated histograms from the sample data we can see that the sample mean fatality rate of the non-bicycle collisions appears to be bigger than the bicycle-only collisions, since the data is concentrated around 1.4 rate, where the motorcycle-only collisions is concentrated around 0.5 rate. Non-bicycle collision data sample appears to be right skewed . Computed sample proportion of



fatal injuries of non-bicycle collisions is: *0.129* and sample proportion of fatal injuries involving bicycles: *0.055*.

Note for visualisation: since data points can depict only 1 collision that can be either fatal or not, data-points will be summarized through grouping 25 data points into 1 and fatality rate for the new data point will be computed. EDA R-code is located in the .rmd file at ***Appendix A***.

Statistical inference using hypothesis testing: We will use statistical inference using hypothesis testing to compare the population proportions between two populations. In our case, the first population (*A*) will be a population of non-bicycle collisions, and the second population (*B*) will define a population of bicycle collisions. Detailed calculations can be found in ***Appendix B***.

Based on the performed statistical inference, under the assumption that there is no difference between two proportions and observed difference of 0.129, we get the p-value (probability of obtaining the observed results) of zero. Concluding that if the true proportion of fatal injuries in the collisions involving bicycles does not differ from the true proportion of fatal injuries in the collisions not involving bicycles, then it's nearly impossible to get an observed difference in fatality rates between the two groups of 0.129 or more. So our first methodology does not support the hypothesis that there is no difference between two proportions.

Bootstrapped confidence intervals methodology: We will use bootstrapped confidence intervals, to compute the probability of observing the sample difference by using the sample data-points. Sample size of around 17000 data-points should be big enough to perform a proper bootstrap sampling. With $\alpha = 5\%$ (significance level), we will construct a 95% interval assuming that both groups' probability distributions are the same in every possible way except perhaps the mean since data points are independent and binomial (Fatal/Non Fatal). Detailed code for bootstrapping and computation of confidence interval can be found in ***Appendix C***.

After completing the simulation under the assumption that there is no difference between 2 proportions, we get the confidence interval of (0.057, 0.097). It implies that if the probability of true proportion of fatal injuries in the collisions involving bicycles does not differ from the true proportion of fatal injuries of observing the difference, based on our sample data, the probability of observing the difference of 0.097 or more is less than 2.5%, so the probability of getting an observed sample difference in fatality rates of 0.129 or more is very small. Hence, data sample does not support the hypothesis that there is no difference between two proportions

---

Conclusion: Based on the performed statistical analysis, we can conclude that data does not support the hypothesis that bicycles are as safe as other vehicles. In the first methodology, we showed that based on the data set used, the probability that the proportions of fatal injuries of both groups are the same is almost 0, and in the second methodology- by using the simulation and confidence intervals, we have concluded that under the assumption that there is no difference the probability of the observing the difference of 0.129 between proportions is very small. It's important to note that both of our methodologies are based on the computed statistics from the sample data (MVC dataset), even though the sample size is pretty big (more than 16 000 data

points), we can not be sure that it's a proper representation of the population (all collision that ever happened in Toronto). As a result we can not answer the research question definitively, but what we can conclude is that our dataset strongly suggests that there is a difference in the true proportions of fatal injuries between the bicycle-involved collisions and collisions not involving bicycles.

References :

[1] Toronto Police . "Motor Vehicle Collisions Involving Killed or Seriously Injured Persons." Toronto Police Services, 2021. May 5. https://data.torontopolice.on.ca/datasets/ksi/explore?location=43.686565%2C-79.406503%2C11.55.

[2] The Netherlands: Ministry of Transport. "Cycling in the Netherlands". Public Works and Water Management. 14 May 2009, http://www.fietsberaad.nl/library/repository/bestanden/CyclingintheNetherlands2009.pdf

MVC Data Analysis

Artem Arutyunov

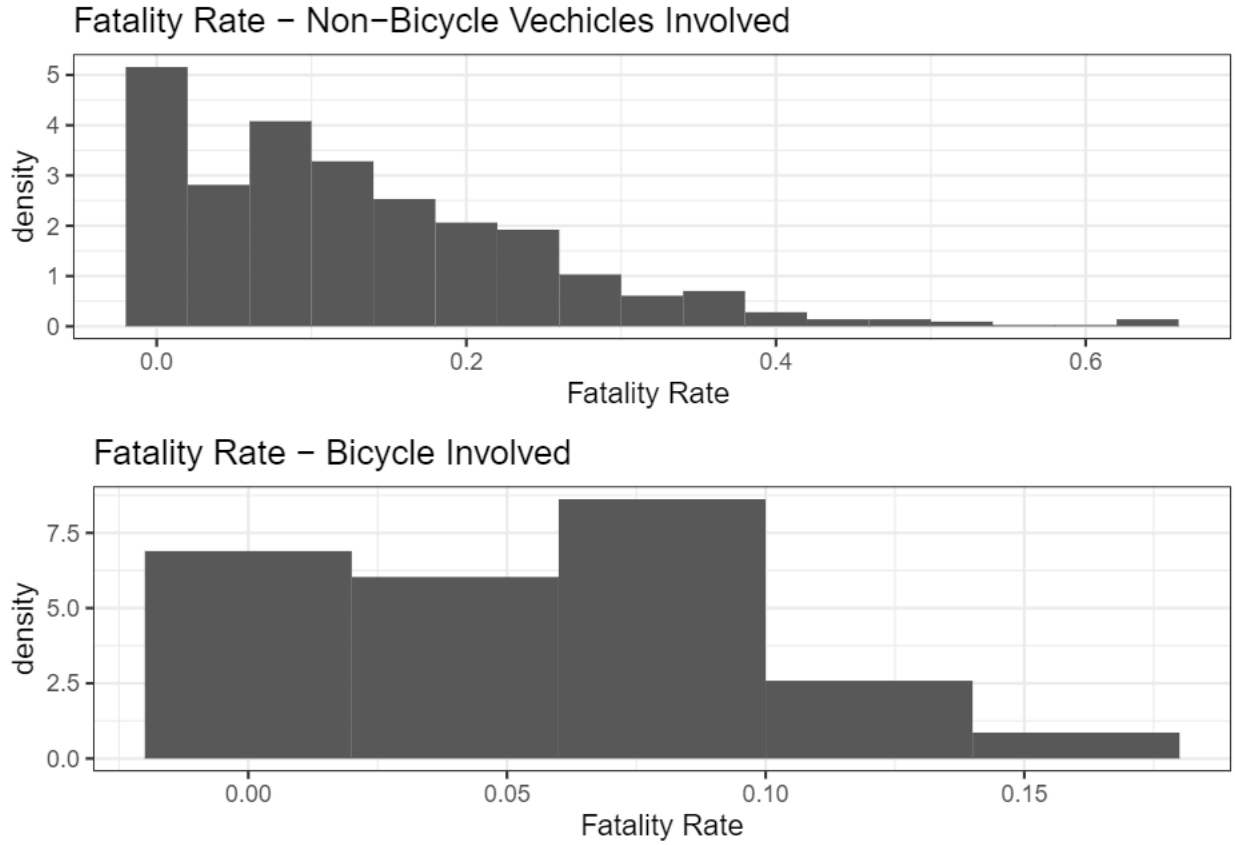
Loading Libraries and Data Cleaning:

```
#Loading the libraries
library(tidyverse)
library(latex2exp)
library(lubridate)
library(gridExtra)

#Loading the data
data <- read_csv('MVC.csv')

#N/A values removed for the needed variables
data <- data[complete.cases(data$ACCLASS),]
data <- data[complete.cases(data$VEHTYPE),]

#Converting from Qualitative to Quantitative
data$ACCLASS <- ifelse(data$ACCLASS == "Fatal", 1, 0)
```

Exploratory Data Analysis:

```
data1<-data[-which(data$VEHTYPE == "Bicycle"), ]

#Computing the summarized statistic per 25 collision each of all-vehicles collisions.
data1His<- data1 %>%
group_by(x = ceiling(row_number()/25)) %>%
 summarize(mean = mean(ACCLASS))


#histogram created for all-vehicles collisions group.
a <- ggplot(data1His)+
  geom_histogram(aes(x= mean, y=..density..),
                 binwidth = 0.04)+
  labs(x = "Fatality Rate ",
       y = "density", title = "Fatality Rate - Non-Bicycle Vechicles Involved")+
  theme_bw()

#Computing the summarized statistic per 25 collision each
data2 <- data[complete.cases(data$VEHTYPE),]

data2<-data2[-which(data2$VEHTYPE !="Bicycle"), ]

#Computing the summarized statistic per 25 collision each
data2His<- data2 %>%
```

```
  group_by(x = ceiling(row_number()/25)) %>%
  summarize(mean = mean(ACCLASS))

#histogram created for bicycle involved collisions group.
b<-ggplot(data2His)+
  geom_histogram(aes(x= mean, y=..density..),
                 binwidth = 0.04)+
  labs(x = "Fatality Rate",
       y = "density", title = "Fatality Rate - Bicycle Involved")+
  theme_bw()

#Producing the histograms
grid.arrange(a,b)
```




```
#Computing the sample fatality proportion of non-bicycle collisions
cat(mean(data1$ACCLASS))
```

```
## 0.1292308
```

```
#Computing the sample fatality proportion of bicycle involved collisions
cat(mean(data2$ACCLASS))
```

```
## 0.05540166
```

Appendix B - Calculations for Methoogy #1:

The sample proportion of fatal injuries of non-bicycle collisions is $\hat{p}_A = 0.129$ and sample size $n_A = 13325$ with 1772 fatal collisions and the sample proportion of fatal injuries of bicycle collisions is $\hat{p}_B = 0.055$ and sample size $n_A = 722$ with 40 fatal collisions. Since we are compairing two proportions we will state the following hypothesis:

$$H_0 : p_A - p_B = 0 \text{ and } H_A : p_A - p_B \neq 0$$

Note that our sample size are large enough since $n_A\hat{p}_A = 1772 \geq 10$, $n_A(1 - \hat{p}_A) = 11553 \geq 110$ and $n_B(\hat{p}_B) = 40 \geq 10$, $n_B(1 - \hat{p}_B) = 682 \geq 10$.

We will have to compute the pooled proportion:

$$\hat{p} = \frac{40 + 1772}{13325 + 722} = 0.129$$

The observation difference between sample proportions is

$$0.129 - 0.055 = 0.074$$

By null hypothesis, we will assume that $p_A = p_B = p = 0.129$, so under null hypothesis we can conduce that:

$$\hat{p}_A - \hat{p}_B \tilde{\sim} N(0, 0.129(0.871)(\frac{1}{722} + \frac{1}{13325})) = N(0, 0.00016)$$

Computing our test statistic we get that:

$$z = \frac{0.074 - 0}{\sqrt{0.00016}} = 5.850$$

Then computing the p-vlues we get that the the probability of $P(|z| \geq 5.850) \doteq 0$.

Appendix C - Bootsrapping code for Methoogy #2:

```
#Setting the indexes
index.bicycle <- data$VEHTYPE == "Bicycle"
obs.bicycle <- data$ACCLASS[index.bicycle]
obs.other  <- data$ACCLASS[!index.bicycle]


#Defining the bootstrap experiment
B <- 2000
boot.mean.diff <- c()

#Performing the sampling simulation and computing the proportion difference
for(b in 1:B){
  boot.bicycle <- sample(obs.bicycle, replace = TRUE)
  boot.other <- sample(obs.other, replace = TRUE)
  boot.mean.diff[b] <-  mean(boot.other) - mean(boot.bicycle)
}

#Computing the confidence interval
ci.mean.diff <- quantile(boot.mean.diff, probs = c(0.025, 0.975))
ci.mean.diff
```

```
##       2.5%      97.5%
## 0.05587057 0.09027947
```