# Citibike miniproject

ChunChiehTsai[1]

[1]New York University (NYU)

November 7, 2017

**Abstract**

I analyzed the behavior of people who ride citibikes based on gender. The gender is a control variable for the trip duration. In trip duration, I found that the mean of trip duration of male is far more than the trip duration of female in January in 2015.

**Introduction**

Citi bike is a public bicycle sharing system in New York city, Jersey City and New Jersey. Also, the Citi bike provides trip histories data for developers, engineers and analysts to find out insights in data. In citibikes open data, I am curious about what is the difference of behavior between male and female. It is interesting for me to discover the difference of trip duration between gender, because I think female tend to be walk or take metro rather than riding bikes. I assume male would ride longer in each ride than the female does. In this hypothesis, the Null hypothesis will be the mean of trip duration of each female is more or the same of the mean of trip duration of male.

**Data**

The data comes from the citibike open data platform. I choose the Citi Bike Trip Histories in January in 2015 as data set. This data set includes trip duration in seconds, start & stop time and date, the name of start and stop station, station id, bike id, user type, user gender and year of birth. (Fig. 1 Data Frame)

| | tripduration | starttime | stoptime | start station id | start station name | start station latitude | start station longitude | end station id | end station name | end station latitude | end station longitude | bikeid |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 634 | 2013-07-01 00:00:00 | 2013-07-01 00:10:34 | 164 | E 47 St & 2 Ave | 40.753231 | -73.970325 | 504 | 1 Ave & E 15 St | 40.732219 | -73.981656 | 16950 |
| 1 | 1547 | 2013-07-01 00:00:02 | 2013-07-01 00:25:49 | 388 | W 26 St & 10 Ave | 40.749718 | -74.002950 | 459 | W 20 St & 11 Ave | 40.746745 | -74.007756 | 19816 |
| 2 | 178 | 2013-07-01 00:01:04 | 2013-07-01 00:04:02 | 293 | Lafayette St & E 8 St | 40.730287 | -73.990765 | 237 | E 11 St & 2 Ave | 40.730473 | -73.986724 | 14548 |
| 3 | 1580 | 2013-07-01 00:01:06 | 2013-07-01 00:27:26 | 531 | Forsyth St & Broome St | 40.718939 | -73.992663 | 499 | Broadway & W 60 St | 40.769155 | -73.981918 | 16063 |
| 4 | 757 | 2013-07-01 00:01:10 | 2013-07-01 00:13:47 | 382 | University Pl & E 14 St | 40.734927 | -73.992005 | 410 | Suffolk St & Stanton St | 40.720664 | -73.985180 | 19213 |

Figure 1: Data Frame

**Methodology**

To test the Null hypothesis, I separate the process to 3 parts including clean, visualization and test. First, data cleaning, I download the data from citibike open data platform and only leave the 2 columns, gender and tripduration. Also drop the raw which has the Null.(Fig 2.)

```
df.columns
```

```
Index(['tripduration', 'starttime', 'stoptime', 'start station id',
       'start station name', 'start station latitude',
       'start station longitude', 'end station id', 'end station name',
       'end station latitude', 'end station longitude', 'bikeid', 'usertype',
       'birth year', 'gender', 'duration', 'duration_second', 'Age'],
      dtype='object')
```

```
df.drop(["duration_second", 'starttime', 'stoptime', 'start station id',
       'start station name', 'start station latitude',
       'start station longitude', 'end station id', 'end station name',
       'end station latitude', 'end station longitude', 'bikeid', 'usertype',"duration","bi
       ], axis=1, inplace=True)
```

```
df.head()
```

| | tripduration | gender | Age |
|---|---|---|---|
| **0** | 634 | 0 | 0 |
| **1** | 1547 | 0 | 0 |
| **2** | 178 | 2 | 37 |
| **3** | 1580 | 0 | 0 |
| **4** | 757 | 1 | 31 |

Figure 2: Data Cleaning

Second, visualization, the histogram can show all the trip duration distribution in each gender. Apparently, the distribution of male of each trip duration is more than the female.(Fig. 3)

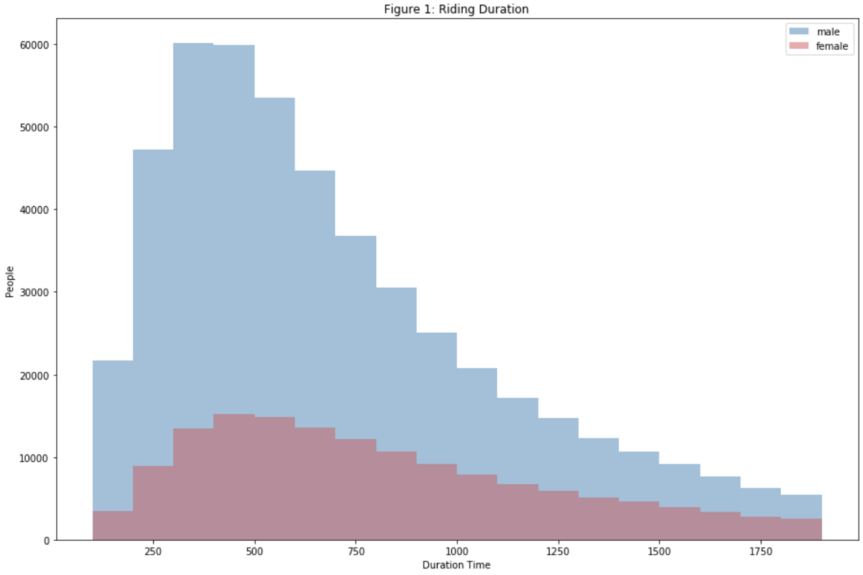


Figure 3: Data Visualization: Duration vs People Count

The last, to determine whether 2 population means are different, I choose z test to address the Null hypothesis. (Figure 3)

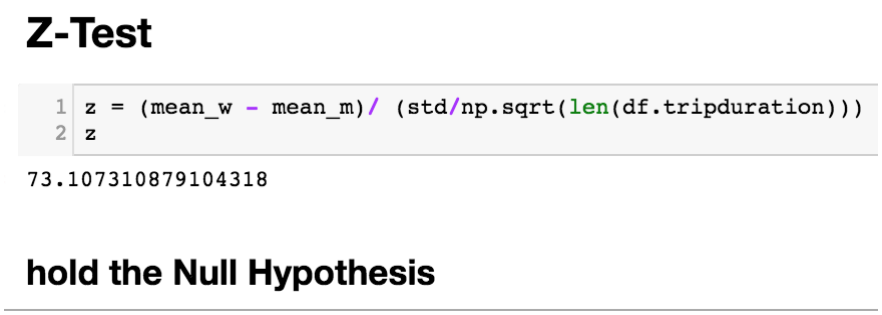


Figure 4: Z Test

**Conclusions**

The p value is 73.11, and it is more than 1.96. This tells us that the Null hypothesis, the female trip duration is more or the same as the male trip duration, is held. (Figure 4)


Figure 5: Average trip duration

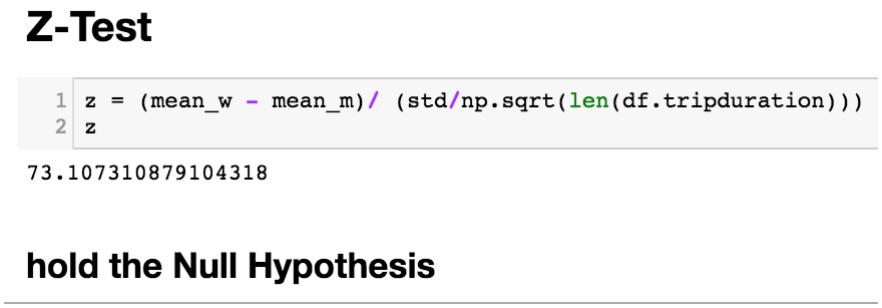


Figure 6: Z test

The strength of this analysis is it can easily to see the results. On the other hand, the weakness of this analysis is it lacks to standardize the data based on the people of each gender. Apparently, male riders are far more than the female riders.