Analysis of Citi-Bike users’ Trip Duration behavior by Gender

sunglyoung, Federica B. Bianco
NYU Center for Urban Science & Progress

INTRODUCTION

IDEA: Women has short trip duration of usage of citibike than men.

NULL HYPOTHESIS:

The women’s average trip duration that use of citi-bike is the same as the men’s trip duration of usage of the bike.

$H_0 : \mu_w = \mu_m$

$H_1 : \mu_w \neq \mu_m$

where μ_w is mean of of Women trip duration, μ_m is mean of Men trip duration

or identically:

$H_0 : \mu_w - \mu_m = 0$

$H_1 : \mu_w - \mu_m \neq 0$

I will use a significance level $\alpha = 0.05$ which means I want the probability of getting a result at least as significant as mine to be less then 5

DATA

The data is selected, from May to July 2017, whole the Citi-bike open data set. Raw data of the Cibi-bike dataset are each month so the three months dataset has merged. The open data contains Trip-duration, start-time, stop-time, 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, user-type, birth year, and gender. Only Trip duration and gender columns are interested in this study so all other columns are dropped. The trip duration is considered only 95% of the total trip duration time, in another word, above 95% of trip duration time is outliners. The Fig.1 shows the higher frequency of male user than the female user. However, total number of male is 3130770 and total number of women is 1087062. The Fig.2 shows normalized of the histogram by area from 0 to 1.

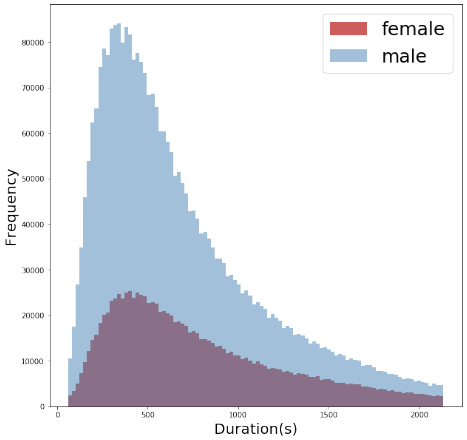

Figure 1. The figure shows male users are higher frequency than female users but the peak trip duration time is closed to each other.

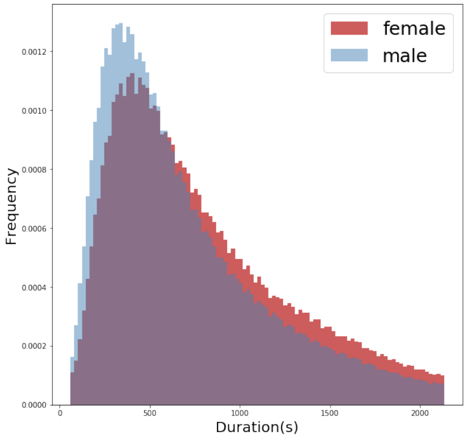

Figure 2. Normalized the data set by dividing total number of each gender. It seems like the two distribution is similar each other but I can’t say before analysis

METHODOLOGY

The null hypothesis is female Citi-bike users’ average trip duration is the same as male Citib-bike users’. The two distribution is not a normal distribution so T-test or Z-test is not compatible to do the hypothesis testing. Therefore, Mann-Whitney U test is chosen to do the test because Mann-Whitney U test is taking the rank of all data by the total number of sample sizes. However, the Mann-Whitney U test requires that the shape of two sample distribution and standard deviation should be similar. female and male standard deviations are 475.89 and 453.02. The shapes look similar to each other. Mann Whitney U test is rank a dataset into from 0 to summation of the number of two sample data size and compare the two distributions median of rank.

Scipy. stats package is used to calculate the p-value by using this code

```
ss.mannwhitneyu(male.tripduration, female.tripduration)
```

The builtin function has a correction to a continuity of function and the result of p-value comes out 0.0.

Conclusion

The Mann Whitney U test shows 1.0 P-value which means female users trip-duration is likely greater than male trip-duration

Another Class mate comments on the hypothesis test. Zhiao mentioned I should not take out outliners, however, the data set have huge standard deviation and cleaned out over 95 % of data is reasonable to me. Also, he suggests using Z-test however, the two distribution is not normal distribution so Z-test doesn't work.