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

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

***Abstract*—Since 2013 the citi-bike has been operated in NYC and the whole dataset has been opened to the public. Analysis citi-bike trip duration behavior by gender by analysis the open data and plotting the two genders trip duration histogram. Null Hypothesis is average of the citi-bike women's trip duration is the same as the men's trip duration. Mann Whitney U test is used to test hypothesis test because the two histogram is non-gaussian distribution. Significant level is chosen to 0.05 and p-value of the hypothesis test is 0.0. Therefore, Null hypothesis can not reject it.**

Introduction

The citi-bike is a bike sharing system that is run by Motivate in NYC, according to Citi-Bike homepage [1]. The system has been operating since May 2013 and it has been operated for 24 hours/day, 7 days/week, 365 days/year because the company installed 600 bike station over 55 neighborhood in NYC [1]. Since the sharing system is embeded on NYC the new type of transportation opens new way of one-way trips, such as commute to school, work, or simple running errands. 10 million trips in one year achieved December 2015 and 100,000th annual member gained on May 2016 [1].

The Hypothesis is women has short trip duration of usage of citi-bike than men and null hypothesis is the women's average trip duration 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. 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, bikei-d, user-type, birth year, and gender. Only Trip duration and gender columns are interested in this study so all other columns are dropped. Above 95% of trip duration time is out-liners and removed from the data set.

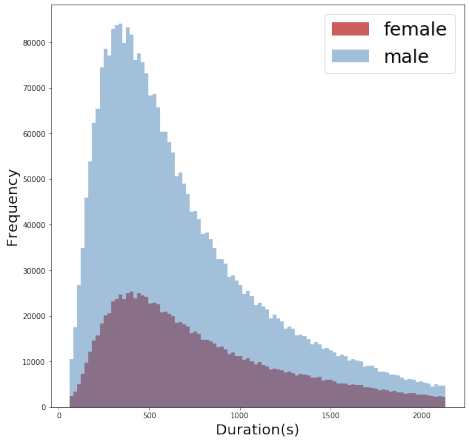

Figure 1. The figure shows male users are higher frequency than female users but the peak trip duration time is closed to each other.

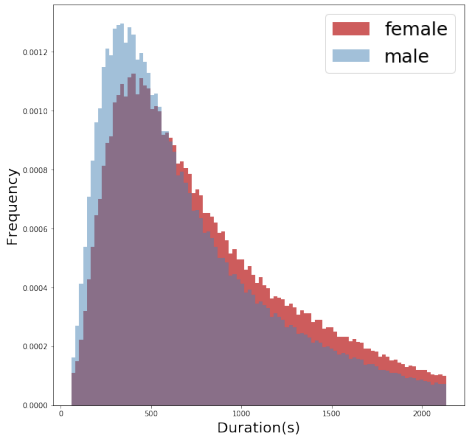

Figure 2. Normalized the data set by dividing total number of each gender. It seems like the two distribution is similar each other but I can't say before analysis

Methodology

The null hypothesis is the average of female Citi-bike users' trip duration is the same as male Citi-bike users. The two

distribution is not a normal distribution so T-test or Z-test is not very compatible to do the hypothesis testing. Therefore, Mann-Whitney U test is chosen to do the test because Mann-Whitney U test can be performed on non-gaussian distributions. The Mann-Whitney U test is taking the rank of all data from 0 to the summation of the total number of sample sizes and comparing between two distribution of median of the ranks.

Scipy.stats built-in function, scipy.stats.mannwhitneyu(x,y), is used to calculate the p-value. The builtin function has a correction to a continuity of function and the result of p-value is 0.0. Therefore, the median of female and male trip duration doesn't have significant differences.

CONCLUSION

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. Federico Bianco commented on that my hypothesis equation was broken and fixed it. The p-value of Mann Whitney U test is 0.0 so the median of female and male trip duration doesn't have significant differences and I can't reject the null hypothesis.

REFERENCES

[1] "About Citi Bike: Company, History, Motivate — Citi Bike NYC," https://www.citibikenyc.com/about, accessed on Tue, November 07, 2017. [Online]. Available: http://www.citibikenyc.com/about