# Who prefer longer citibike riding? Men or Women?

yl5240[1]

[1]Affiliation not available

November 9, 2017

**Abstract**

It is interesting to discuss about the riding characteristic of different gender. One of them is that whether the trip duration of men is significantly different from women. This article use two month citibike data, July 2017 and Dec. 2016, and use t test to check the hypothesis that women prefer longer citibike riding. The conclusion is that the proportion of women riding more than 30 minutes is significantly larger than that of men at the significant level 0.05.

The notebook of analysis: https://github.com/picniclin/PUI2017_yl5240/blob/master/HW7_yl5240/HW7_(HW3_assignment2)_yl5240.ipynb

**Introduction**

Citibike is a pulic bike rent service offered by Citi group. User could buy a card for a day's riding or could subscribe to join as a monthly, quarterly or yearly member. The latter, also called subscriber, could enjoy a much lower charge. So far there are more than 800 citibike stations across manhattan, brooklynn, queens and new Jersey. Users could rent a bike locked on a dock from any station, and ride it to explore the city or to commute. There is time limit for every ride, the first 45 minutes of which is included with membership or in the customer's fee. Overtime fees would be charged if the user did not return bike in time. Since there're a upper limit for regular use, I am wondering is there any difference between men and women on using the public bike.

**Data**

I choose the citibike data of July 2017 and Dec. 2016, one in summer and one in winter. First, I only select two columns, gender and tripduration, in the dataset. Then, I transfer the trip duration of seconds into trip minutes, making it more sense. Third, I group the new data table by trip minutes, dividing the duration into seven parts including 0-10, 10-20, 20-30, 40-50, 50-60, 60-120, to see the whole trend and identify the difference between two gender(See Fig. 1).

**Methodology**

First, I group the dataset again, but only into two groups: more than or fewer than 30 minutes trip duration. 30-minute was selected as the threshold of long trip, considering the 45-minute time limit.

Second, I compare the absolute number(See Fig. 2) as well as the fraction(See Fig. 3) of men and women on riding more than or fewer than 30 minutes. In the absolute number plot, we could see although there are far more men riders than women, the gap shrink on long trip riders. And when we come to the fraction plot, the trend is reversed, and the women gain the upper hand on long-trip percentage.

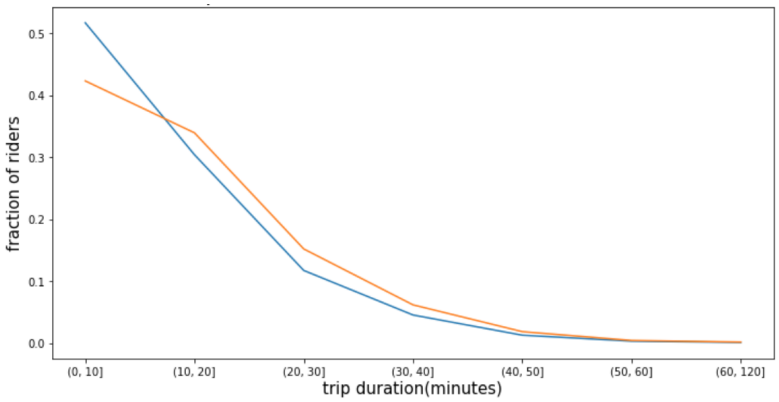


Figure 1: Fraction of men and women in different trip duration section

Last, I use t test to check does the women really prefer longer bike trip. The null hypothesis is that the percentage of women riding more than half an hour is lower than that of male.

As Baoling has suggested, for 2 groups in dataset, we should use 2 groups unpaired t test.

Another choice is z test. Because the observation and degree freedom is large enough, it's good to adopt the z statistics as an alternative approach.

The process of t test is that : 1) get the proportion of long trip for each gender via the dataset, 2) calculate the standard error using the formula(See Fig. 4) for difference between two samples' proportion, 3) check does the proportion of women-long-trip is significantly larger.

For the last step, two approach could be adopted. One is to calculate the z score, and check does it larger than 1.96 (alpha = 0.05). The other is to calculate the lower bound of the confidence interval (alpha = 0.05) of the fraction of women-long-trip, and check does it larger than the the fraction of men's

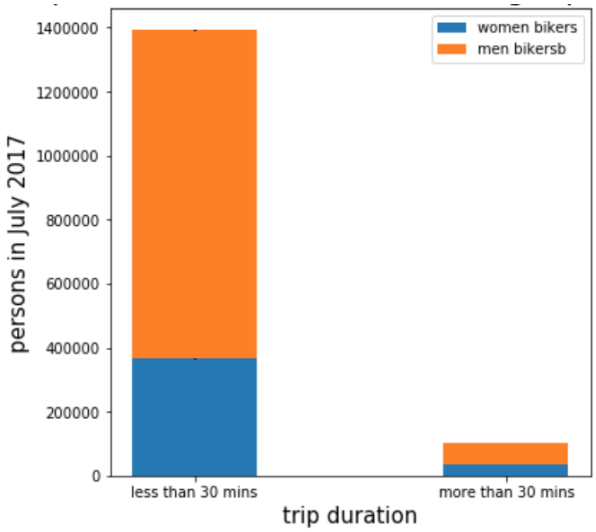


Figure 2: Comparison of men and women long trip duration

**Conclusions**

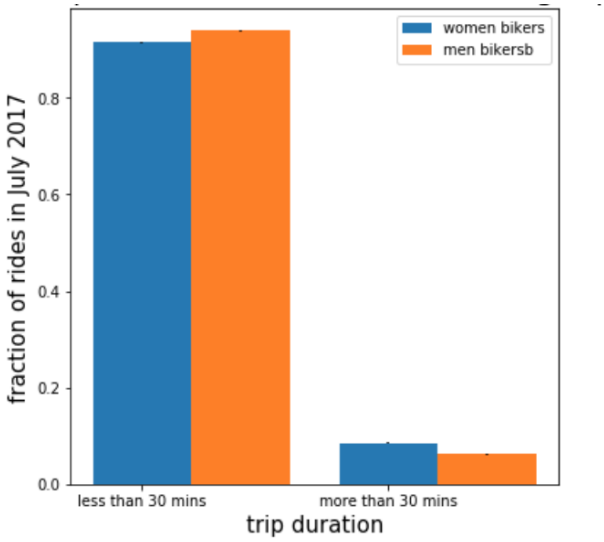


Figure 3: Fraction comparison of men and women long trip duration

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Figure 4: formula of standard error for two samples' proportion difference

The proportion of men riding over 30 minutes is smaller than that of women at the significant level 0.05. The z-score of Jul. 2017 is 47.6 and that of Dec. 2016 is 20.4, both way larger than the 1.96 (alpha = 0.05). And the percentage of men-long-trip in the two month are both significantly lower than the lower bound of 95% confidence interval of the percentage of women-long-trip.

So we reject the null hypothesis that women has fewer percentage of longer trip duration. Both the summer and winter sample prove that women riders prefer longer rides than men. The hypothesis is robust to seasonality.

But the biggest problem about this test is that the t test is parametric and the distributions may not comply to the assumptions because the variables is categorical and not Gaussian distributed. I need to further check does it make sense to use the t test in this way. The reason I didn't adopt the chi square test, usually for two groups' test, is that I am not sure whether Chi Square Goodness of Fit with Null hypothesis of two distributions being the same or Chi Square Test of Independence with Null hypothesis of two categorical variables being independent, is suitable for this topic.