Women actually bike more at morning —- Citibike data analysis

Zhiao Zhou[1]

[1]NYU Center for Urban Science & Progress

November 7, 2017

Citibike Dataset Analysis (For NYU CUSP PUI2017 HW7)

<Zhiao Zhou, zz1749>

Abstract

New York Times has reported that there were more male bike-share members in NYC where about a third of members were female, who cared more about safety and convenience. However, it was also mentioned that quite a few women liked biking to work.(TIMES, 2015) So it would be interesting to find out if the ratio of men biking at morning (commuting period for most people) over man biking the whole day is smaller than the ratio of women which would help balance the gender disparity.

Here we carried out a z test between proportions in iPython notebook to test my hypothesis using a sample of 201706 Citibike (The most popular bike-sharing system in NYC) public datasets. It turned out that the Z-score is 9.9977 and the p-value is 7.7958e-24. So we could accept our alternative hypothesis that women actually bike more at morning which would be useful for future analysis since the existing gender disparity seems to result from lack of infrastructure and safety for women.

Introduction

First launched in 2013, Citibike has now totals of 706 stations and 12,000 bikes which pushed itself to become the biggest bike-sharing system in the USA.(wik) However, Citibike has been struggling to figure out why men far outnumber women in using their services, with the number of men riders double that of women riders, as Sarah M. Kaufman, the assistant director of tech programming at the Rudin Center for Transportation at NYU, said that women became early indicators of a successful bike system which means that if you had more women riders, it means that it would be convenient and safe. (FITZSIMMONS, 2015) This phenomenon also emerged in Chicago and Washington where bike-sharing systems attracted more men. And till now it's still not solved yet what triggers this gender disparity. The Citibike company was trying to introduce new stylish bikes or add new stations to woo women.

What was fun was that there seemed to be a number of women who loved to commute by public-sharing bike. If we could find out that in fact, women bike more than men at morning, the company could focus more on service for women during a commute. Additionally, this hypothesis was untested, we could easily test it using z-test, nonetheless. Figure 1 shows my null hypothesis and its corresponding maths expression as well as my significance level.

After doing the z-test for two proportions, we found that our hypothesis was justified. So regardless of men riders outnumbering women riders, women are more willing to bike at morning, which Citibike company

Null Hypothesis:

The ratio of man biking at morning (5 am to 12 am as the interval between sunrise and noon) over man biking the whole day is the same or higher than the ratio of woman biking at morning over woman biking the whole day

$$H_0 : \frac{W_{\text{morning}}}{W_{\text{day}}} <= \frac{M_{\text{morning}}}{M_{\text{day}}} \quad ¶$$

$$H_1 : \frac{W_{\text{morning}}}{W_{\text{day}}} > \frac{M_{\text{morning}}}{M_{\text{day}}}$$

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%

Figure 1: Null hypothesis and the significance level chosen

could attach importance to if they tried to lure more women subscribers.

Data

The datasets used for the test was 201706 Citibike dataset within the CUSP data facility (DF), which we could easily access working on NYU CUSP compute platform. Then I used iPython notebook to process my datasets (you could access the original notebook file click on the upper left side of Figure 2).

```
In [17]: data.columns

Out[17]: Index([u'tripduration', u'starttime', u'stoptime', u'start station id',
       u'start station name', u'start station latitude',
       u'start station longitude', u'end station id', u'end station name',
       u'end station latitude', u'end station longitude', u'bikeid',
       u'usertype', u'birth year', u'gender'],
      dtype='object')
```

Figure 2: The columns of the original datasets

The original datasets contained a lot of datasets as in Fig. 2which we didn't need so I dropped most of them and kept only two of them that needed as in Fig. 3in that we were trying to analyze the fraction of the number of rides at morning based on gender.

```
In [18]: data = data[['gender','starttime']]

In [19]: data.head()
```

Out[19]:

	gender	starttime
0	1	2017-06-01 00:00:02
1	1	2017-06-01 00:00:13
2	1	2017-06-01 00:00:20
3	2	2017-06-01 00:00:24
4	1	2017-06-01 00:00:33

Figure 3: Dropping useless columns

Now that the starttime column was in the type of string which would be hard to process so we converted it into datetime type which contains methods for users to simply extract hour, minute, seconds and so on as in Fig. 4. What's more, I changed the name of that column into "date" for clarification.

```
In [20]: data= data.rename(columns={'starttime':'date'})
         data['date'] = pd.to_datetime(data['date'])
         data.head()
```

Out[20]:

	gender	date
0	1	2017-06-01 00:00:02
1	1	2017-06-01 00:00:13
2	1	2017-06-01 00:00:20
3	2	2017-06-01 00:00:24
4	1	2017-06-01 00:00:33

Figure 4: Converting string of dates into datetime

Till now, the datasets were finally cleaned so that we could move to our methodology.

Methodology

First, we visualized the distribution of Citibike users' fraction of the frequency of ridings on 24 hours a day by gender in order to first have a glimpse if our hypothesis would make sense from a plot. Then we got Fig. 5 which showed us that from 5:00 to 12:00 it seemed that women tend to bike more thus our hypothesis made sense and we could do a statistical test now, which was also why I finally determined the time period of a morning between 7:00 to 12:00 as was questionned in the peer review.

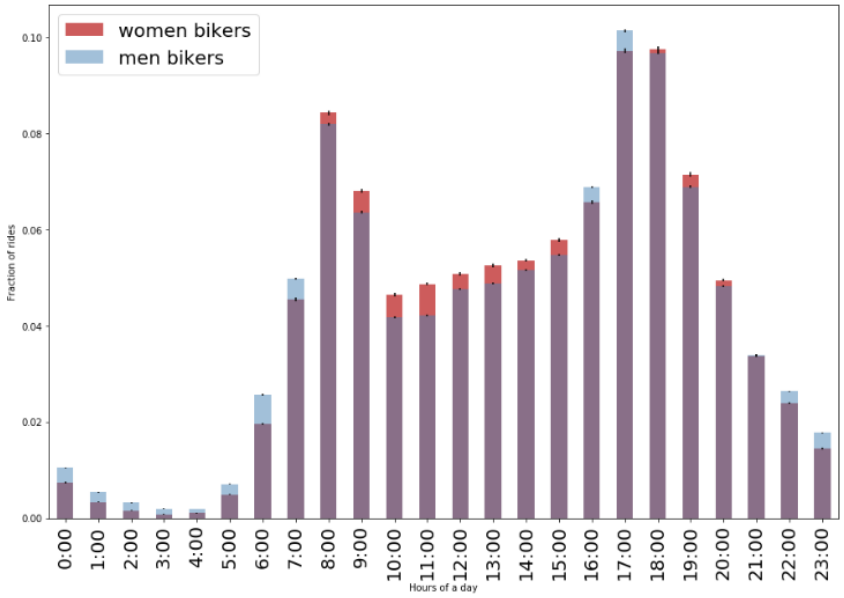

Figure 5: Distribution of Citibike users' frequency of rides by gender in June 2017, normalized

Here we could use both parametric or non-parametric tests. Thanks to my peer Unisse Chua (uc288) 's review, I found her recommendation very useful and resonable. Since the sample came from the same population of Citibike users with paired data and there were at least 30 observations per sample, we could apply a z-test for two proportions to test the hypothesis. However, now that the comparison given was between ratios and a priori expectation is being done, based on the chart from the slides, the test to be used could also be the chi-squared goodness of fit test with Yate's correlation or Fischer's exact test. And for a chi-square test for equality of two proportions is exactly the same thing as a Z-test and due to function convenience in iPython, I was using a z-test.

First, I calculated the number of rides at morning and the total number of rides a whole day based on different genders using following codes as in Fig. 6(same as men):

counts_w = data.date[data.gender == 2].groupby([data.date.dt.hour]).count()

norm_w = counts_w.sum()

w_morning = counts_w.loc[5:12].sum()

```
In [24]: w_morning = counts_w.loc[5:12].sum()
         p_w = w_morning/norm_w

In [25]: m_morning = counts_m.loc[5:12].sum()
         p_m = m_morning/norm_m

In [26]: print(p_m, p_w)

0.359593791135 0.368420921464
```

Figure 6: Calculation of necessary arguments for the test

Then I used the **proportions_ztest** method imported from **statsmodels.stats.proportion** which could output the z statistics and p value after inputing the counts and total numbers of two samples as follows where value means the difference between the proportions and alternative means if I want to apply a two-sided test or one-sided one.

stat, pval = proportions_ztest(counts, nobs,value=0, alternative='larger')

Conclusions

```
counts = np.array([w_morning, m_morning])
nobs = np.array([norm_w, norm_m])
stat, pval = proportions_ztest(counts, nobs, value=0, alternative='larger')
print('z-score is {}, p-value is {}'.format(stat, pval))

z-score is 9.99773865967, p-value is 7.79583622734e-24
```

Figure 7: Result of the Z-test

We could see from Fig. 7 that p-value is way smaller than 1 - our significance level —- 5%, so we could confidently reject our null hypothesis and conclude that the ratio of men biking at morning over man biking the whole day is smaller than the ratio of woman biking at morning over woman biking the whole day. This was really an unexpected outcome since people always think that men bike more than women do.

The Z-test led to a mighty credible outcome at last. The Citibike company could take into consideration in the future to add more stations in office areas or reduce the rent rate at morning and the reason for this

phenomenon might also be women are more concerned about safety issue after the morning in that road harassment or reckless driving are still kind of common in the city.

There was as well some weakness in this mini-project that needed to be solved in future studies. First, the sample was only one-month datasets which could have seasonal effects. Second, the sample was still not big enough. But this is now still a good gap filling.

References

Citi Bike - Wikipedia. https://en.wikipedia.org/wiki/Citi$_B ike.URL$. Accessed on Tue, November 07, 2017.

EMMA G. FITZSIMMONS. A Mission for Citi Bike: Recruiting More Female Cyclists. *THE NEW YORK TIMES*, Page A16, 2015. URL https://www.nytimes.com/2015/07/08/nyregion/a-mission-for-citi-bike-recruiting-more-female-cyclists.html. Accessed on Tue, November 07, 2017.

THE NEW YORK TIMES. Times Readers React to Citi Bike's Gender Gap. *THE NEW YORK TIMES*, Page A16, 2015. URL https://www.nytimes.com/2015/07/11/nyregion/times-readers-react-to-citi-bikes-gender-gap.html. Accessed on Tue, November 07, 2017.