Are younger people more likely to use CitiBikes than older people?

Sean Andrew Chen[1]

[1]Affiliation not available

November 9, 2017

Abstract - Cycling is a great alternative for urban mobility needs. It is cheap, sustainable, and does not contribute to congestion as much as other modes of transport. Despite these great benefits to urban life, cycling may still not meet the needs of the entire population equitably. Since cycling requires physical exertion, it may not be an adequate mode of transport for a significant part of the population. Thus, we test if the proportion of CitiBike users who are under 50 is greater or equal to the proportion of users over the age of 50. We find using a z-test that indeed the proportion of users under 50 is larger, accounting for a whopping 80% of the system.

Introduction - CitiBike is New York City's bikeshare system. Such a system allows users to rent bicycles for a short amount of time, picking them up from one station and depositing them at an end destination station. Bikeshare systems have become more and more popular across the globe in major world cities as it allows a cheap, sustainable, and healthy alternative for transportation. However, for the system to be truly effective, it must reach all segments of the population. Thus, it is prudent to know which segments of the population the system is currently or is not currently serving. Our thesis is that because of the nature of cycling as a physical activity, those with lesser physical abilities will use the system less. If we hold the hypothesis that as one ages, one's physical abilities decline, then we would expect to see that the system is heavily biased towards younger riders. That is, we hypothesize that the proportion of system users that are "young" greatly outweighs the proportion of users that are "old". It is important now to define what we mean by old or young. We use a study from the *Journal on Gerontology* which finds that a person's physical abilities begin to decline in their 50s. Thus we will use 50 as a cutting off point, creating a binary category in the data for young or not.

If we find this hypothesis to be true, this gives impetus to the argument that the system should perhaps have an electrified component giving riders with diminished physical abilities a boost up and opening the system up to them to use as well, not just the young and healthy. Urban transportation is not just about efficacy but also equity. What use is a system if it does not serve the entire population equitably? We ensure that our subway stations have elevators as part of compliance with the Americans with Disabilities Act. It is of course true that to use a cycling system, one must have a baseline level of physical abilities and it would be nonsense to get rid of the entire system because it does not meet the needs of those with more severe mobility challenges. But it is nevertheless important to try to open the system up to as many people as possible by identifying the challenges that they currently face. By investigating any difference in the age of users and by hypothesizing a link to physical abilities attached to age, we can perhaps identify one of these challenges and perhaps a way to remove this barrier.

Data - CitiBike data is offered open source providing information on user birth year, user type (subscriber to the system or one off user), trip start time and end time, start station ID and end station ID. We used data from two years 2015 and 2016 in the months of January and August. We used two years to smooth out the data to account for unobservables year to year. We used these two months to smooth out seasonal differences as well. However, these are two of the most extreme months. A further study would perhaps try to use samples from all months of the year. We did not do this simply because of computational limitations and time limitations.

Once downloading the data, we cleaned it up, removing extraneous variables to enhance the speed of computation and creating an age variable from birth year by subtracting the birth year from the year the data was created. By creating a histogram of the age of users in year, we can see already that the distribution is almost Poissonian, heavily favoring younger users, peaking around 30 and declining from thereon.

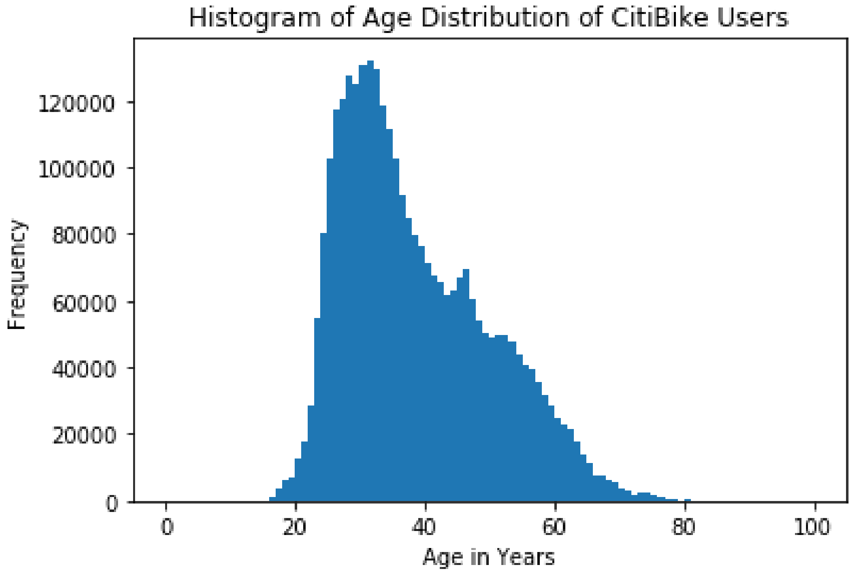

Figure 1: Histogram of age distribution of CitiBike Users

Methodology -

To see if the proportion of users under 50 is greater than or equal to users over the age of 50, we first find those proportions. We find that it is roughly an 80%/20% split.

From there, we perform a one tailed Z-test to test equality of proportions. Because the null hypothesis stipulates that each proportion would be 50% - as this is a binary categorical variable - we use 50% as the original proportion from which to subtract the proportion of young users. Ultimately, we find the z statistic to be well above the 1.96 threshold for a level of 95% confidence, meaning we can reject the null hypothesis.

The reviewer of my original proposal recommended the use of the t-test. While the t-test is indeed very handy, I believe that the very large n (sample size) allows us to forgo the use of the t-test. Moreover, the t-test is an updated version of the z-test for tests of low n. Thus, the z-test should still work. However, a Chi Square test should be the better fit nonetheless due to the categorical nature of the data that we have created - user age group being a binary.

We did perform such a Chi Square test using different years as extra categorical variables with the idea of seeing if this proportion held up over different years. However, the test resulted in a very high p-value.

Conclusions -

Our data was limited to only two different months of two different years. Ultimately, we would like to retest this with a larger diversity of time to account for any unobservables. Also the 50 years of age cutoff while

$$z = \frac{\hat{\pi} - \pi_0}{se_0}$$

$$se_0 = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$$

Figure 2: Z Test for Proportions

not necessarily arbitrary creates only a binary variable. It would be more helpful to create multiple age bins, especially seeing as how the modal years of age was around 30.

Nonetheless, since we can conclude that indeed the number of users under age 50 is statistically significantly larger than those over, we can be assured that we have found a demographic weakness in the bikeshare system: it is not attracting an entire age demographic. We hypothesis this is due to physical abilities. If that hypothesis holds true, then things like electrification with electric motors could be a way to remove this barrier and open up the system to that demographic. However, that hypothesis itself still needs to be tested. We do not fully know if it is because of physical fitness. It could be because of certain prejudices and mental heuristics held by that age demographic - e.g., cycling to them is seen as a young person thing to do or is thought to be too dangerous. This will need further investigation.

Figure 3: To launch Jupyter Notebook, click where it says Code to the left.