Assignment 2-Hw8

Borong Lyu[1]

[1]New York University

November 4, 2018

Abstract

This paper tested the null hypothesis that The average trip duration of men is equal to that of women. The alternative hypothesis is that: The average trip duration of men is different from that of women. This paper uses a significance level α=0.05.

This paper used unpaired 2 sample T-test to test this hypothesis. Before applying T-test. This paper used Fisher's F test to test the hypothesis of equality of the variances. Then the result shows that these two samples have unequal sample sizes and unequal variances.

As a result, the T statistics is smaller than the critical value. Thus, this paper rejects the null hypothesis. The trip duration of the female is significantly different from the trip duration of the male.

Introduction

Citi Bike is the nation's largest bike share program, with 12,000 bikes and 750 stations across Manhattan, Brooklyn, Queens, and Jersey City. It was designed for quick trips with convenience in mind. Citybike saves money and time. Citybike is fun to ride and is a good way to exercise. The New York CityBike has been widely used by New Yorkers. Thus, the study of CityBike data is deep and profound.

(According to cb4184's review, I had my hypotheses backward. I think He/She is right about that. Great thanks to cb4184. In order to simplify my following hypothesis testing. I changed my null hypothesis. However, I think since we don't know anything about the whole population. An unpaired 2 sample T-test is the appropriate test here. Not Z-test.)

Data

Used the City Bike API to download data. The dataset includes the 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, bike id, user type, birth year, and the gender of riders in January, 2016. Firstly, we dropped the columns that are not needed. Only trip duration and gender are needed in this hypothesis. Then, we grouped the data by gender. The data set is now classified into two categories: 1, the trip duration of women and 2, the trip duration of men.

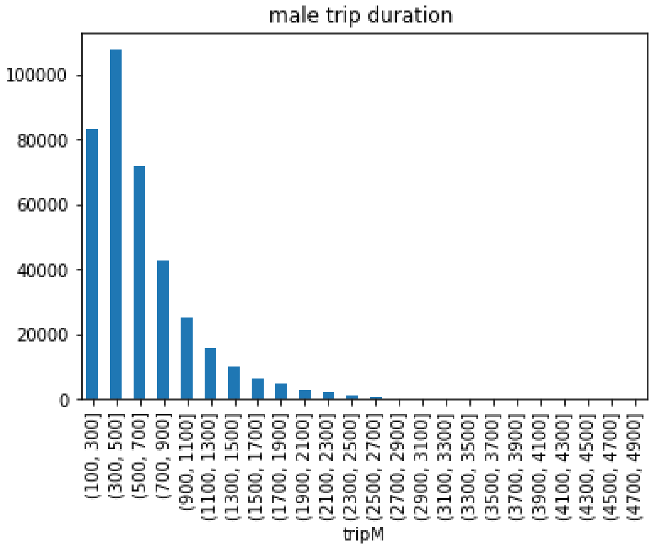

Figure 1: The bar plot of the distribution of male riders' trip duration

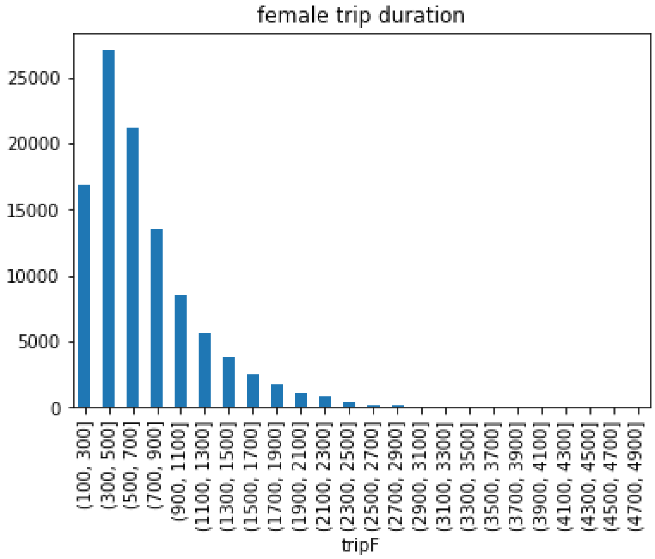

Figure 2: The bar plot of the distribution of female riders' trip duration

Figure 1 and Figure 2 are the distribution of riders' trip duration by gender. The shape of these two distributions seems similar. Slight difference could be spotted: The trip distribution of Female seems higher than that of Male.

Methodology

Take a sample S1(The trip of men) comprising $n1=379314$ observations, of mean $\mu1= 728.169$ and standard deviation $s1 = 5251.629$.

Take a second sample S2(The trip of Momen), comprising $n2=104457$ observations, of mean $\mu2=900$ and variance $s2=9751.866$.

.This paper adopts two-sample unpaired t-test. Since the true variance of the populations from which the samples are extracted is unknown;

$n_1 = 378912$. $n_2 = 104283$.

The use of Student's t-test requires a decision to be taken beforehand on whether variances of the samples are to be considered equal or not. This paper uses Fisher's F test to test the hypothesis of equality of the variances and to use the result of the test in the subsequent calculations.

The F hypothesis test is defined as:

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 \neq \sigma_2^2$

where

$F = \frac{s_1^2}{s_2^2}$

S_1^2

and

S_2^2

are the sample variances. Here this paper uses the significance level of 0.05. The statistic is 0.29. The Critical values are : $F(1-\alpha/2,N1-1,N2-1) > 0.79$ and

$F(\alpha/2,N1-1,N2-1) < 1.03$. The F test indicates that there is enough evidence to reject the null hypothesis that the two batch variances are equal at the 0.05 significance level.

So the variances are different. If we consider that the variances are different, the statistic is given by:

$t = (\mu1 - \mu2 - D) / ([?]s1^2/n1 + s2^2/n2)$

$t = -5.4922$

The T statistic follows a Student distribution with $n1+n2-2 = 483769$ degrees of freedom.

According to the T-table, the critical value for the two-tailed t-test is +/- 1.96.

Conclusions

As a result, the T statistic is smaller than the critical value. Thus, we reject the null hypothesis. The trip duration of the female is significantly different from the trip duration of the male. (The trip duration of women is, in fact, longer than that of men.)