Hypothesis Testing

Linfeng Zhou

July 17, 2015

Question 1

Statement #1: The majority of the trips are for short commutes lasting no more than 15min.

- What should be your null hypothesis? The null hypothesis is:

$$H_0 : \mu \leq 15min$$

- What is a reasonable alternative?

$$H_1 : \mu > 15min$$

- What type I error test are your conducting?
  The type I error I am conducting is 0.05.
- What is significant level in your test?
  The significant level in my test is 0.05.

Question 2

Statement #2: Citi Bike System wants to tackle bike rides incurring in overtime fees, particularly their interest is in rides lasting more than 45min.

- Test the hypothesis that the Median for overtime is 2hrs long with alpha=5% (overtime = 45mins or more).

  (1) Method I: t(student) test

  According to requirement, the null hypothesis and alternative hypothesis are:

$$H_0 : \mu > 120min \Leftrightarrow H_1 : \mu \leq 120min$$

  μ is population mean, the sample mean is: $\overline{X}$=86.6032407. Using pre-built `t.test` function, the details of t-test are following:

```
t.test(subsetdata2-120)
```

```
##
##  One Sample t-test
##
## data:  subsetdata2 - 120
## t = -5.8809, df = 215, p-value = 1.545e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -44.59013 -22.20339
## sample estimates:
## mean of x
## -33.39676
```

(2) Method II: Sign test

T-test's result may have error since the test dataset are required two properties:

1. The dataset should be samll sample.
2. The sample are from norm distribution.

Due to those reasons, we can use other statistical diagnostic methods to decrease error. One of those methods is sign test which is a non-paramatric hypothesis test model. It is more robust than t-test when we deal with asymmetric distribution.

The null hypothesis and alternative hypothesis are:

$$H_0 : M_e > 120min \Leftrightarrow H_1 : M_e \leq 120min$$

M_e is median of dataset.

```
binom.test(sum(subsetdata2<120),length(subsetdata2),0.5)
```

```
##
##  Exact binomial test
##
## data:  sum(subsetdata2 < 120) and length(subsetdata2)
## number of successes = 184, number of trials = 216, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.7973424 0.8964075
## sample estimates:
## probability of success
##              0.8518519
```

- Provide a paragraph discussing the findings and statistical significance of the test.

  Both p-value of t-test and sign test are smaller than 0.05, therefore we reject the null hypothesis which assumed that the median for overtime is 2hrs long with alpha=5%.

Question 3

Citi Bike management thinks that men incur in more overtime fees. Test this hypothesis by comparing overtime variances across genders.The null hypothesis and alternative hypothesis are:

$$H_0 : \sigma_1^2 = \sigma_2^2 \Leftrightarrow H_1 : \sigma_1^2 \neq \sigma_2^2$$

```
table(gender)
```

```
## gender
##   1   2
## 154  62
```

```
aggregate(tripduration,by=list(gender),FUN=mean)
```

```
##   Group.1        x
## 1       1 87.27284
## 2       2 84.94005
```

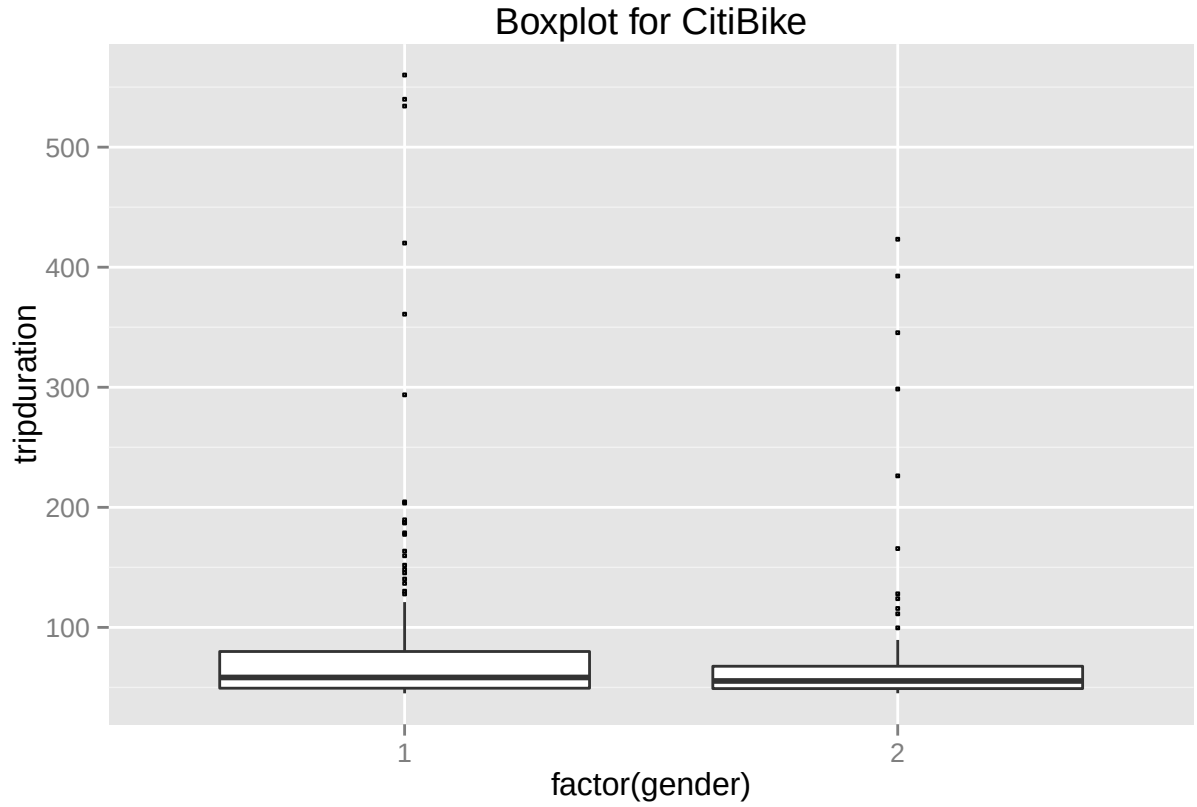
```r
aggregate(tripduration,by=list(gender),FUN=sd)
```

```
##   Group.1        x
## 1       1 84.56717
## 2       2 81.30382
```

```r
summary(aov(cleandata$tripduration ~ cleandata$gender))
```

```
##                      Df  Sum Sq Mean Sq F value Pr(>F)
## cleandata$gender      1   23711   23711   121.5 <2e-16 ***
## Residuals         20851 4068784     195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the means, it appears that trip duration of men are greater than women is an overtime situation. We can observe that the f-statist is 121.5 with a p-value of 2e-16 which means the ANOVA F test for gender is significant. We clearly reject the null hypothesis of equal means for both gender.



From the boxplot it appears that the mean trip duration for gender =”1” is greater than that of gender “2”.