UCSL: R Challenge 2

Christopher Prince (cmp670@nyu.edu)

07/22/2015

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

1.1 Test plan

We first state the null and alternative hypotheses. Denoting the population median with $\tilde{\mu}$,

$$H_0 : \tilde{\mu} \leq 900$$

$$H_a : \tilde{\mu} > 900$$

Because the alternative hypothesis is a "greater than" condition, we need to use an upper-tail hypothesis test. For a t-test, this means that we will reject the null hypothesis if the statistic is greater than the critical value for the test at the desired confidence. We will assume $\alpha = 0.05$.

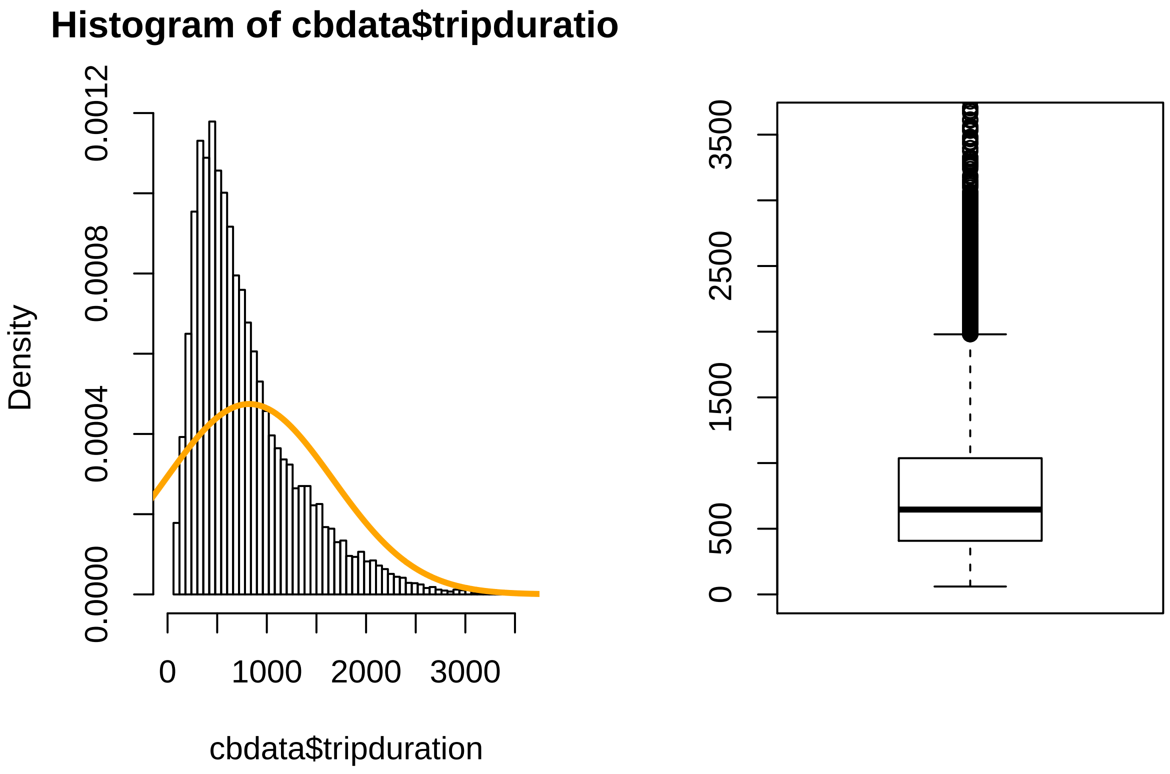
1.2 Importing and summarizing the data

```r
#import Jan 2015 'clean' CitiBike data from csv file, include a header line
fname <- '~/UCSL/R/ch2//Citi Bike Clean Data.csv'
cbdata <- read.csv(fname, header = TRUE)
```

```r
#Print the summary data and calculate the sample parameters n, s, xbar and xmed:
#kable(head(summary(cbdata$tripduration)))

n <- dim(cbdata)[1]
s <- sd(cbdata$tripduration)
xbar <- mean(cbdata$tripduration)
xmed <- median(cbdata$tripduration)
```

```r
#Fit a normal distribution to the data. Overlay a normal distribution curve onto the histogram of the d
par(mfrow = c(1,2))
hist(cbdata$tripduration, breaks = c((1:60)*60,Inf), probability = TRUE, xlim=c(0,3600))
xfit <- seq(xbar-4*s,xbar+4*s)
yfit <- dnorm(xfit, mean = xbar, sd = s)
lines(xfit, yfit, lwd=3, col='orange')
boxplot(cbdata$tripduration, ylim=c(0,3600))
```

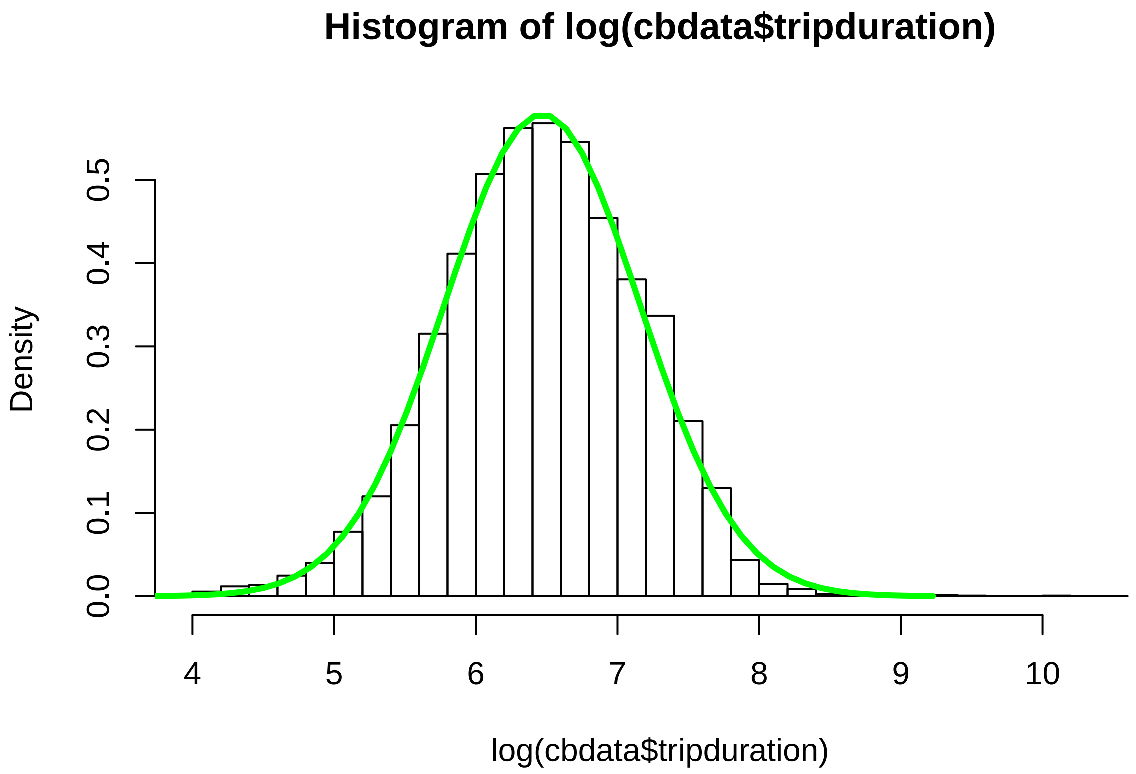


1.3 Determining an analysis method

Before we even do any analysis, we note that the null hypothesis is heavily favored in this sampling, since 68.2% of values are less than 900. It would be remarkable for a sample of over 20,000 observations be so off-balance from a true median over 900. Nevertheless, we perform the calculations to support this.

Because $n = 20853$, and the histogram is not wildly skewed (out-of-normal), CLT applies. But how do we the hypotheses using t-tests when they are on medians and not on means? CLT tells us that means of a sufficiently large number of samples from a population of any distribution will themselves distribute normally. Does this apply to medians as well? Intuitively, subsampled medians seem like they should distrubute normally in the same manner that the mean does. Let's run an experiment. First, the histogram suggests that this may follow something like a log-normal distribution. The histogram of the log of the `tripduration`s looks like so:

```
xbarlog<-mean(log(cbdata$tripduration))
slog<- sd(log(cbdata$tripduration))
xfitlog<-seq(xbarlog-4*slog,xbarlog+4*slog, length.out = 50)
yfitlog<-dnorm(xfitlog,mean=xbarlog,sd=slog)
hist(log(cbdata$tripduration), breaks=30, probability = TRUE)
lines(xfitlog,yfitlog,lwd=3,col='green')
```

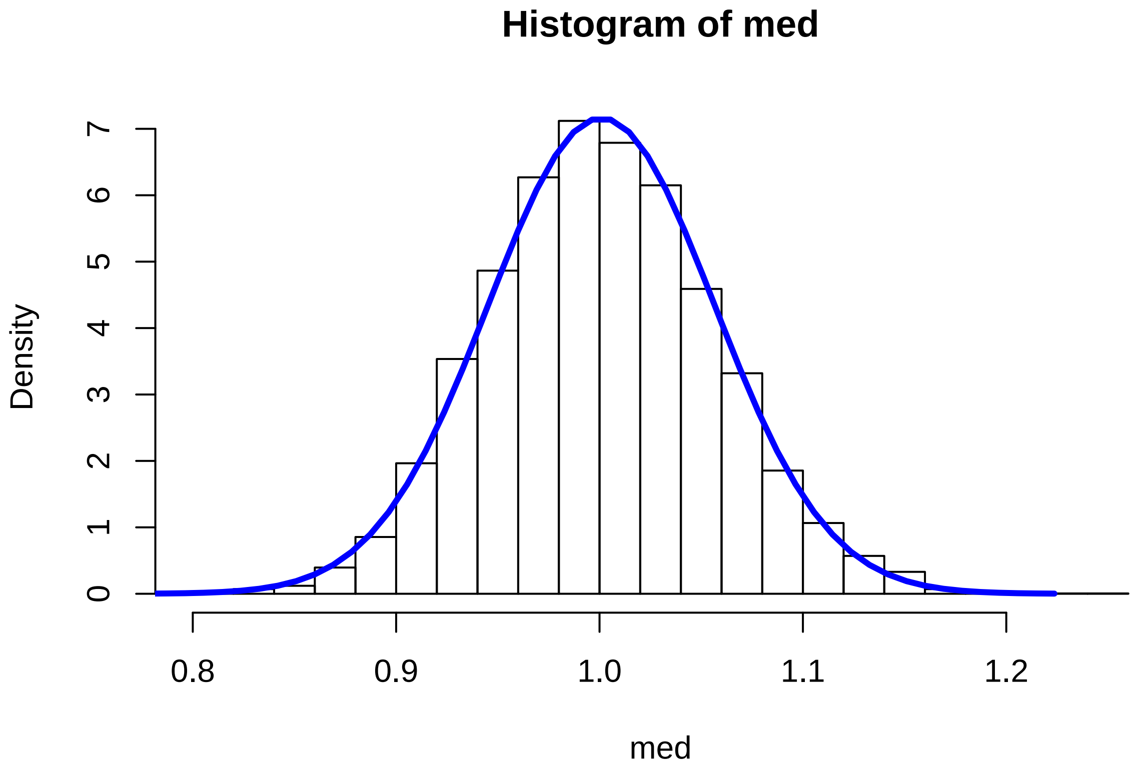


Looks convincing! It's certainly close enough for us to use the lognormal distribution as a model for the experiment. We will create 10,000 random samples of the lognormal distribution and record their medians:

```
med<-array()
for(i in 1:10000){
  m<-rlnorm(n=500)
  med[i]<-median(m)
}
```

Plotting the histogram of the medians and overlaying a normal fit:

```
hist(med, breaks=30, probability = TRUE)
medmean<-mean(med)
meds<-sd(med)
xfitmed<-seq(medmean-4*meds,medmean+4*meds, length.out = 50)
yfitmed<-dnorm(xfitmed, mean=medmean, sd=meds)
lines(xfitmed, yfitmed, lwd=3, col='blue')
```



Based on this analysis, the sample median distribution is normal, so we can be confident using a t-test for the median. Since the population variance is not known we will use the t-statistic instead of the z-statistic. We calualate the t-statistic and we will reject the null hypothesis if it is greater than the critical value. Instead of testing the location of the mean, we test on the location of the median:

$$t = \frac{\tilde{x} - \tilde{\mu}}{s/\sqrt{n}}$$

The only remaining uncertainty is how to adjust the standard deviation s.[1] But as we shall see below, any such correction would have to be enormous to have an impact on the outcome of the test.

```
med_h0 <- 900   #the hypothesized median
alpha <- 0.05 #significance level

t1<-(xmed-med_h0)/(s/sqrt(n))
t1
```

```
## [1] -43.63616
```

```
qt(1-alpha,df=n-1)
```

```
## [1] 1.644927
```

[1] The value for the variance of a sample median cited in various online sources is

$$s^2 = \frac{1}{4nf(\tilde{\mu})^2}$$

.

```r
#Or, the p-value:
pt(t1,df = n-1)
```

```
## [1] 0
```

Because the t-statistic is not greater than the critical value–by quite a large margin–we do not reject the null hypothesis.

1.4 Using an alternative test

The Wilcoxon signed-rank test is used explicitly to test for the median location of a sample.

```r
wilcox.test(cbdata$tripduration, mu=900, alternative = 'g')
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  cbdata$tripduration
## V = 72074000, p-value = 1
## alternative hypothesis: true location is greater than 900
```

Here, we the p-value must be less than α to reject the null hypothesis. Clearly, it is not, so we do not reject it.

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

Subset the overtime data:

```r
cbd_ot<-subset(cbdata, tripduration>2700)
med_ot<-median(cbd_ot$tripduration)
```

$$H_0 : \tilde{\mu}_{OT} = 7200$$

$$H_a : \tilde{\mu}_{OT} \neq 7200$$

Since the null hypothesis tests for an equality, we need a two-tailed t-test. We reject the hypothesis if the t-statistic is larger than the critical value (or is less than the negative critical value). Using the same value for α, we have:

```r
med_ot_h0 <- 7200
alpha <- 0.05
s_ot <- sd(cbd_ot$tripduration)
n_ot <- dim(cbd_ot)[1]

t2 <- (med_ot-med_ot_h0)/(s_ot/sqrt(n_ot))
t2
```

```
## [1] -11.03948
```

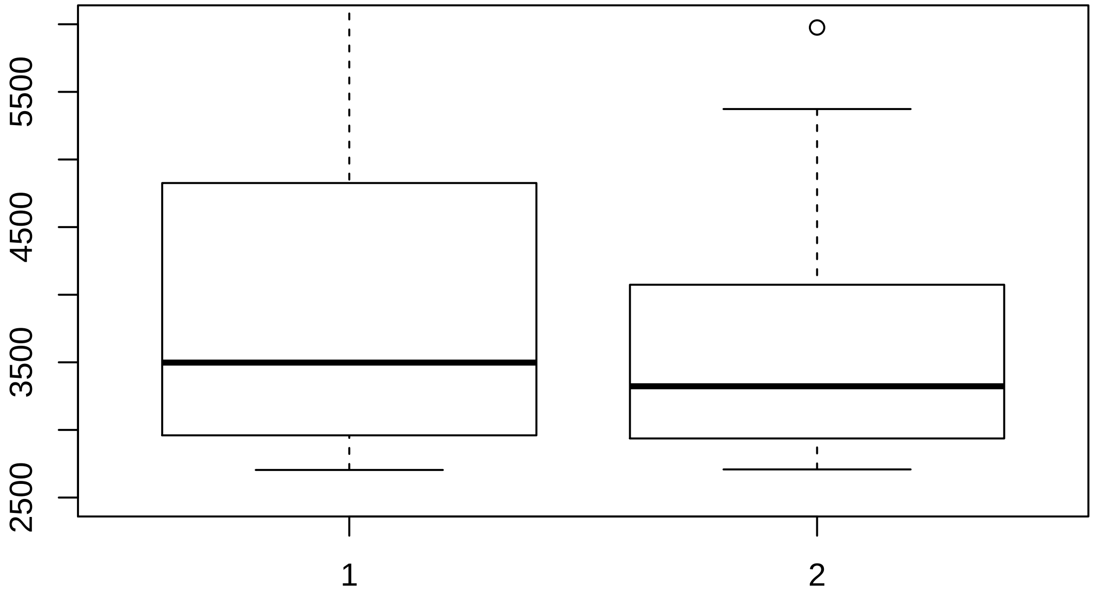
```
qt(1-alpha/2,df=n_ot-1)
```

```
## [1] 1.971059
```

Since the t-statistic is less than the negative critical value, we reject the null hypothesis that the median for overtime is 2 hours.

3 *Citi Bike management thinks that men incur in more overtime fees. Test this hypothesis by comparing overtime variances across genders.*

```
boxplot(tripduration ~ gender, data=cbd_ot, ylim=c(2500,6000))
```



$$H_0 : \text{no difference in OT mean by gender}$$

$$H_a : \text{difference in OT mean by gender}$$

This is exactly the problem posed in the text of this week's lesson, but for the subset of overtime data. Thus, we use ANOVA with the same value of $\alpha = 0.05$:

```
anova <- aov(tripduration ~ gender, data = cbd_ot)
summary(anova)
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## gender        1 8.660e+05   865983   0.034  0.853
## Residuals   214 5.391e+09 25190322
```

The critical value for the test is:

```
qf(1-alpha,1,n_ot)
```

```
## [1] 3.88487
```

Since the F statistic is not greater than the critical value, we do not reject the null hypothesis. The P-value of 0.853 also indicates that we should not reject the null hypothesis with $\alpha = 0.05$