STAT 350
Lab7
Tian Qiu
Part A

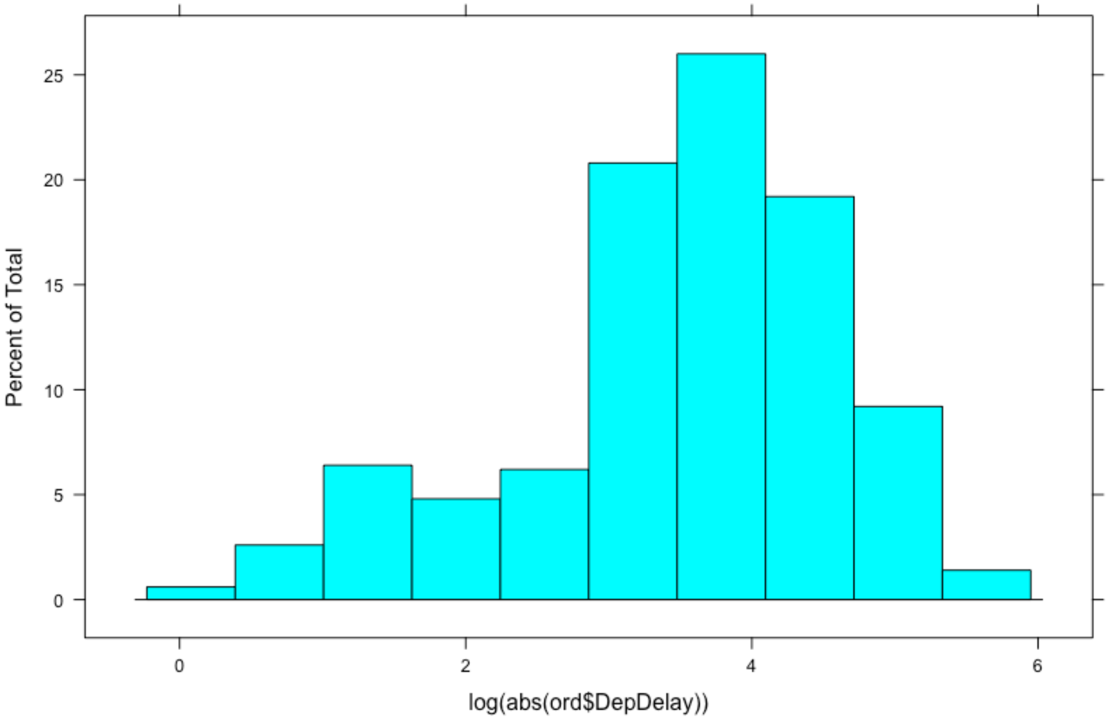
```
1. setwd("~/Desktop/Purdue/STAT350_R/STAT350/Labs/Lab7")
CHICAGO=read.table(file="airline_cleaned.txt",header=T)
studynew <- subset(CHICAGO, Origin=="ORD" | Origin == "MDW")

attach(studynew)
# Note: it is required that you have two curves (red and blue) on the
# histogram and the line on the normal quantile plot as done by the
#     code below.
library(lattice)
histogram(~log(abs(DepDelay)) | Origin, layout=c(1,2),type="density",
panel=function(x)
{panel.histogram(x)
   panel.mathdensity(dmath=dnorm,col="blue",lwd=2,args=list(mean=mean(x,
na.rm=T), sd = sd(x,na.rm=T)))
   panel.densityplot(x,col="red",lwd=2)
   })
bwplot(~log(abs(DepDelay)) | Origin, layout = c(1, 2), pch = "|") #Boxplots side-by-side
qqmath(~log(abs(DepDelay))| Origin, data = studynew, panel = function(x){
   panel.qqmath(x)
   panel.qqmathline(x)
})

# t test:
#t.test (qual ~ categories,conf.level=C, mu = mu0, paired=FALSE,
#alternative="value", var.equal = FALSE)
#is used for confidence intervals and hypothesis tests
#the qualitative variable is first, the variable with the groups in it is second.
#The difference is in first alphabetically – second alphabetically
#   conf.level = C = 1 - alpha
#   for the hypothesis test. mu is mu_0
#   paired = FALSE (2 - sample independent)
#alternative = "greater" or "less" or "two.sided" (this is the appropriate alternative
hypothesis)
#var.equal = FALSE (the variances are not equal, R calls
#                              the Satterthwaite approximation the Welch approximation)
```

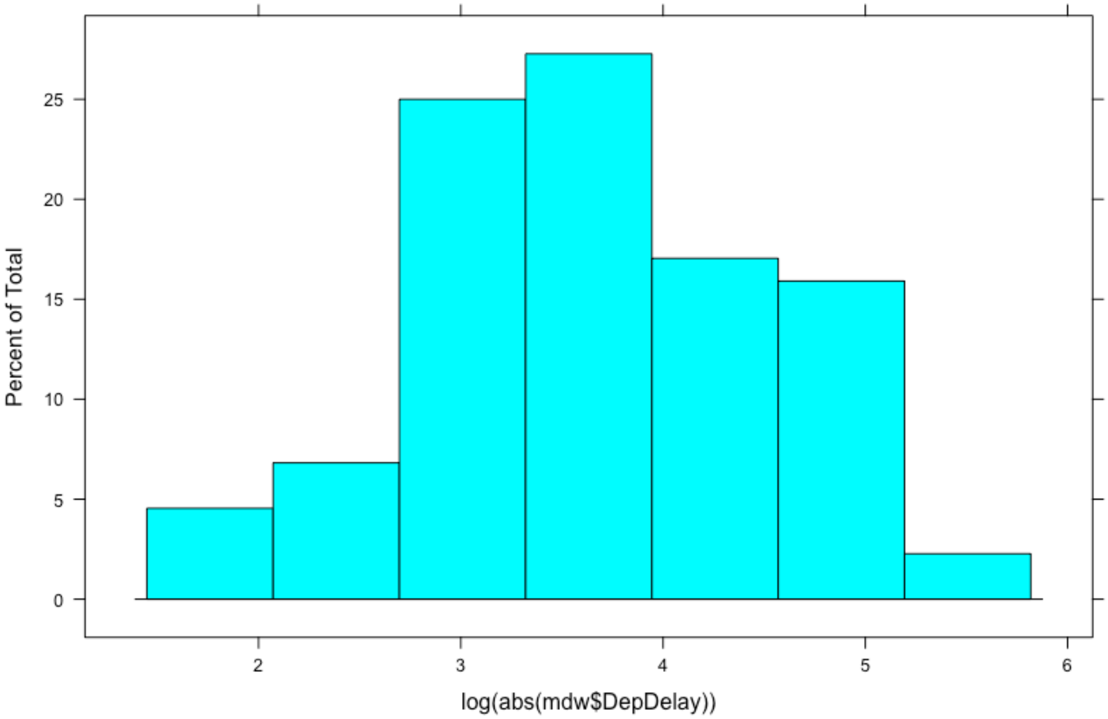
```
t.test(DepDelay ~ Origin, studynew, mu=0, conf.level=0.95,paired=FALSE,
alternative = "two.sided",var.equal=F)
# Information required for f
stdmen = sd(subset(studynew,Origin == "ORD")$DepDelay)
sizemen = length(subset(studynew,Origin == "ORD")$DepDelay)
stdwomen = sd(subset(studynew,Origin == "MDW")$DepDelay)
sizewomen = length(subset(studynew,Origin == "MDW")$DepDelay)
se = sqrt(stdmen^2/sizemen + stdwomen^2/sizewomen)
```

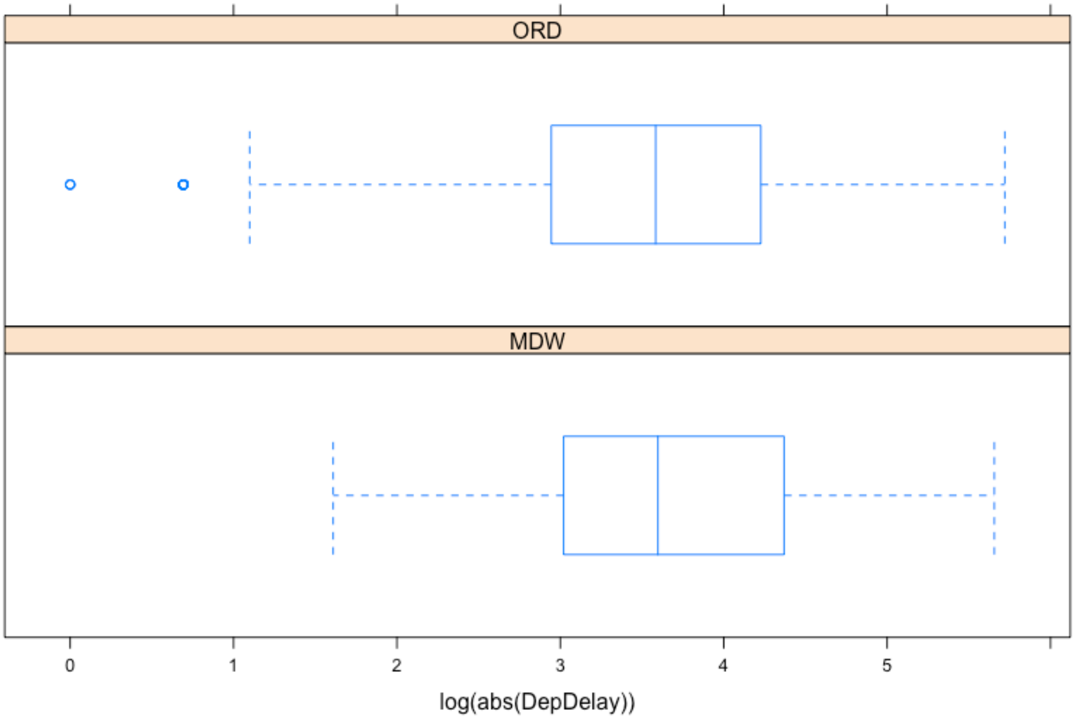
2. We should use two sample independent procedure to analyze the data because these two set of data come from completely two sources that are each independent.

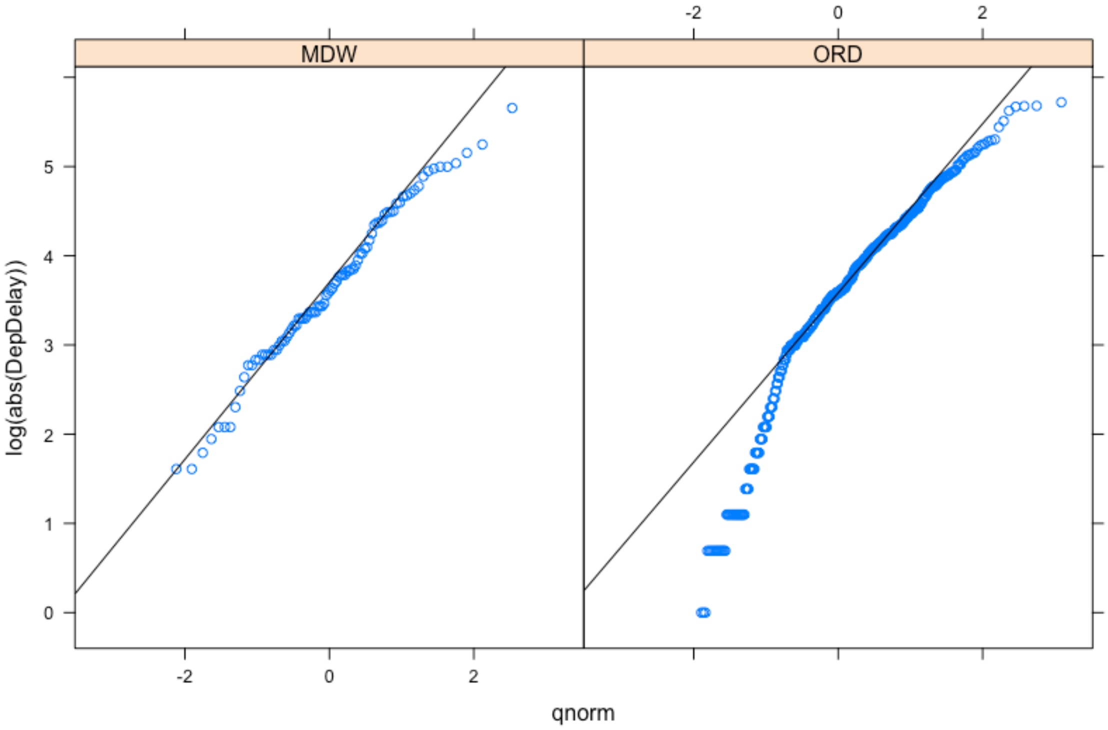
3. We should use two tails because we are comparing the difference and we are not sure which one is bigger.

4. From the graph below we can see that the data from midway is quite normal while the data from ORD has a few outliers on the left side.




Percent of Total
0
5
10
15
20
25
2
3
4
5
6
log(abs(mdw$DepDelay))


ORD
MDW
0
1
2
3
4
5
log(abs(DepDelay))



5. 95 percent confidence interval:

(-7.226076 15.917349)

We are 95% confident that the difference between transformed time of ORD and MDW is between -7.226076 and 15.917348.

6. Ho : μ0-μd = 0 Ha: μ0-μd ≠ 0

t = 0.74362, df = 118.81, p-value = 0.4586 > 0.05

Hence we fail to reject the null hypothesis.

7. They are actually stating the same thing. Our 95% confidence interval contains 0 means that we fail to reject the difference is 0 at significance level = 0.05.

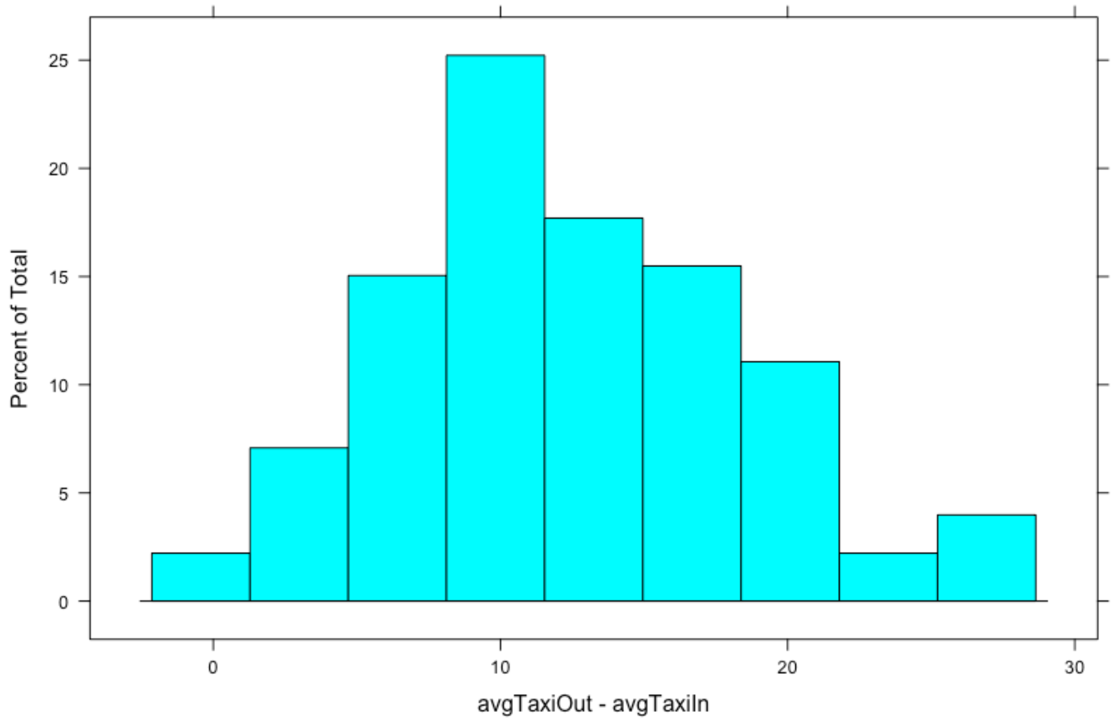
Part B

```
1. setwd("~/Desktop/Purdue/STAT350_R/STAT350/Labs/Lab7")
mpg=read.table(file="AirlineTaxi_In_Out.txt",header=T)
attach (mpg)
# For the diagnoistics plots, you will need to create the one sample #
data which is the difference between the two sets. You will need to #
create the histogram, boxplot and QQPlot on this data set
# (code not included)
normaltest = avgTaxiOut - avgTaxiIn
histogram(avgTaxiOut - avgTaxiIn)
#t.test (x,y,conf.level=C, mu = mu0, paired=TRUE, alternative="value")
# is used for confidence intervals and hypothesis tests
# conf.level = C = 1 - alpha
# for the hypothesis test. mu is mu_0
#   paired = TRUE (2-sample paired)
# The pairing will be x ??? y
# alternative = "greater" or "less" or "two.sided" (this is the
#    for two-sample independent ONLY appropriate alternative
hypothesis)
t.test(avgTaxiOut,avgTaxiIn, mu=0,conf.level=0.99,paired = TRUE,
alternative = "two.sided")
t.test(avgTaxiOut,avgTaxiIn, mu=10,conf.level=0.99,paired = TRUE,
alternative = "greater")
# Information required for f
std = sd(normaltest)
size = length(avgTaxiOut)
se = std/sqrt(size)
```

2. We should use 2-sample pairs procedure because we are using the data from same sets of taxis.

3. I would consider using one-sided because it is already being indicated that the time arriving at the gate is less than the time that go out from gate.

4.



From the histogram we can see that it is normally distributed.

5. 99 percent confidence interval:

(11.29968 13.39043)

We are 99% confident that the true time difference between out from gate minus time into the gate is between 11.29968 and 13.39043 minutes.

6. Ho : μd =10 Ha: μd > 10

t = 5.8277, df = 225, p-value = 9.662e-09 < 0.01

Hence we reject the null hypothesis and with a significance evidence the mean difference is larger than 10.

7. They are actually the same because 10 is not in our 99% confidence interval so we reject it.