STAT 579 Homework 3

Yifan Zhu

September 29, 2016

Problem 1

(a) read data

```r
require(gdata)  # load package
wind <- read.xls(xls = "http://maitra.public.iastate.edu/stat579/datasets/wind.xls",
    perl = "C:/Perl64/bin/perl.exe")  # read xls file and assign it to a data frame
```

(b) calculate the summary statistics

```r
summary(wind) # min, Q1, median, Q3, max, mean
```

```
##      Spring          Summer           Autumn         Winter
##  Min.   :  0.0   Min.   : 10.00   Min.   : 30   Min.   : 50.0
##  1st Qu.: 55.0   1st Qu.: 20.00   1st Qu.:155   1st Qu.:205.0
##  Median :185.0   Median : 35.00   Median :215   Median :255.0
##  Mean   :176.7   Mean   : 80.83   Mean   :200   Mean   :238.3
##  3rd Qu.:275.0   3rd Qu.:150.00   3rd Qu.:260   3rd Qu.:297.5
##  Max.   :350.0   Max.   :190.00   Max.   :350   Max.   :340.0
```

```r
sapply(wind, sd) # standard deviation
```

```
##    Spring    Summer    Autumn    Winter
## 123.97458  72.92067  94.00193  86.63752
```

```r
sapply(wind, IQR) # IQR
```

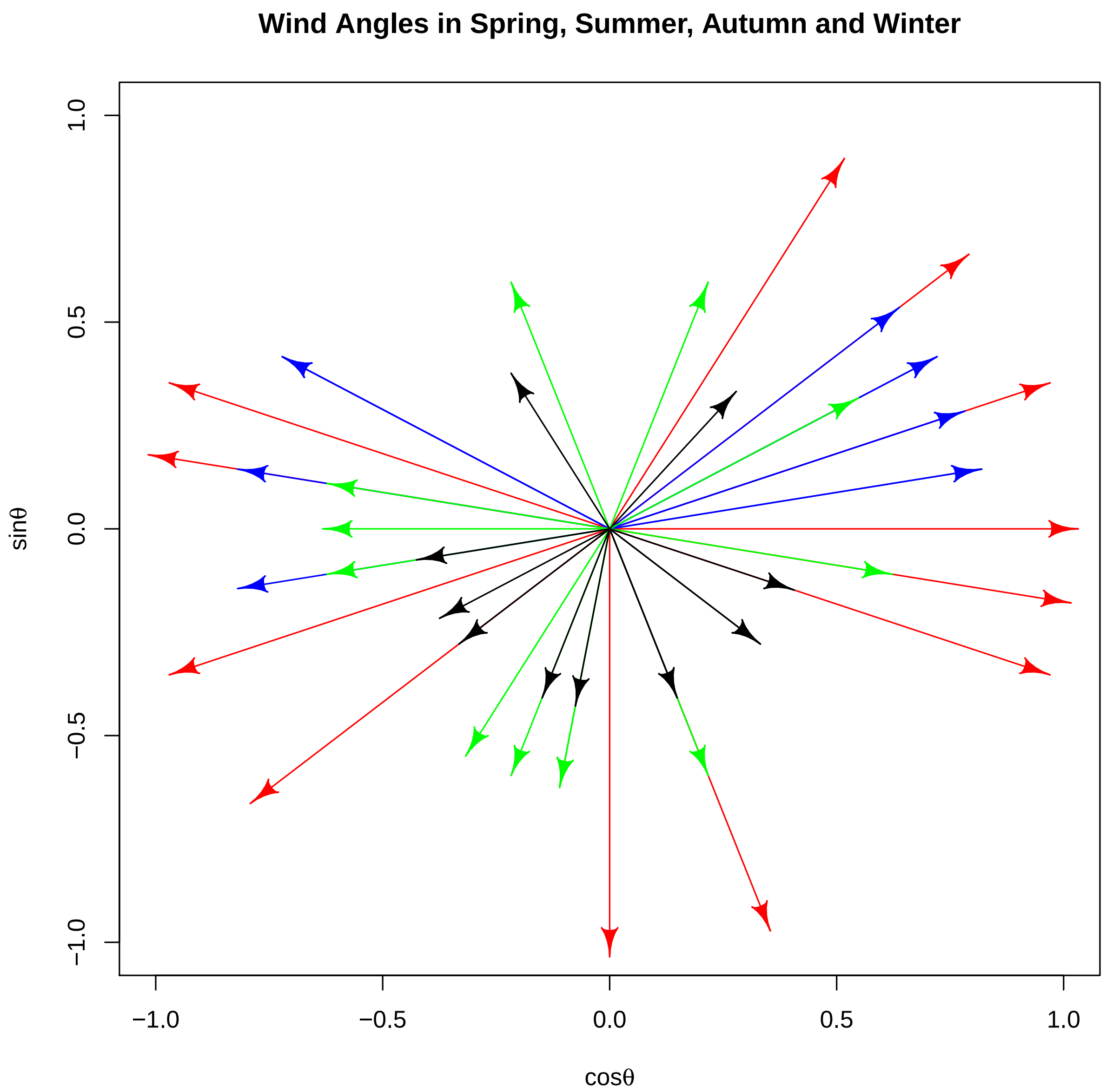
```
## Spring Summer Autumn Winter
##  220.0  130.0  105.0   92.5
```

(c) They all make sense because the values of these statistics are all between 0 to 359.

(d) plot the angles in $(\cos\theta, \sin\theta)$. We use arrows to represent the angle, and use different lengths to avoid overlapping.In the figure, red arrows represent spring wind angles; blue arrows represent summer wind angles; green arrows represent autumn wind angles and black arrows represent winter wind angles.

```r
require(latex2exp)
require(shape)
costheta <- cos(pi * wind/180)
sintheta <- sin(pi * wind/180)
plot(x = -1:1, y = -1:1, "n", xlab = TeX("$\\cos \\theta$"), ylab = TeX("$\\sin \\theta$"),
    main = "Wind Angles in Spring, Summer, Autumn and Winter")
Arrows(x0 = 0, y0 = 0, x1 = costheta$Spring, y1 = sintheta$Spring, col = "red")
```

```
# For angles of wind in summer, the length is reduced to 0.8.
Arrows(x0 = 0, y0 = 0, x1 = 0.8 * costheta$Summer, y1 = 0.8 * sintheta$Summer,
    col = "blue")
Arrows(x0 = 0, y0 = 0, x1 = 0.6 * costheta$Autumn, y1 = 0.6 * sintheta$Autumn,
    col = "green")
Arrows(x0 = 0, y0 = 0, x1 = 0.4 * costheta$Winter, y1 = 0.4 * sintheta$Winter,
    col = "black")
```



In spring, the wind is mostly to west, east and northeast; in summer, the wind is mostly to west and east; in autumn and winter, wind is to all directions.

Problem 2

(a) read data.

```
bikes <- read.csv("C:/Users/fanne/Desktop/STAT579/bikes.csv",
    header = T)
nrow(bikes)
```

```
## [1] 77186
```

```
str(bikes)
```

```
## 'data.frame':    77186 obs. of  8 variables:
##  $ Duration       : int  420 600 1080 540 540 1320 600 1020 300 120 ...
##  $ Start.date     : Factor w/ 8362 levels "6/1/2014 0:00",..: 8098 8098 8097 8097 8097 8097 8097 809
##  $ wday           : Factor w/ 7 levels "Fri","Mon","Sat",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ hour           : int  23 23 23 23 23 23 23 23 23 23 ...
##  $ Start.Station  : Factor w/ 315 levels "10th & E St NW",..: 4 2 72 48 4 153 9 140 225 236 ...
##  $ End.Station    : Factor w/ 315 levels "","10th & E St NW",..: 11 46 104 11 11 38 5 81 205 236 ...
##  $ Subscriber.Type: Factor w/ 2 levels "Casual","Registered": 2 2 2 2 2 2 1 1 2 2 ...
##  $ Bike.          : Factor w/ 2418 levels "","W00007","W00008",..: 774 1777 882 1555 2258 1916 406 4
```

There are 77186 trips. There are 6 factor variables and 2 others.

(b) i.

```
summary(bikes$Duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     420     720    1094    1140  775600
```

```
775600 / (24 * 60 * 60)
```

```
## [1] 8.976852
```

The longest rental is 9 days.

ii. The rental time varies a lot. There is a big spread.

iii.

```
# count the number of trips whose duration was longer than one day
sum(bikes$Duration > (24 * 60 * 60))
```

```
## [1] 7
```

7 retals lasted for more than one day.

(c) i.

```r
nstart <- table(bikes$Start.Station)
# find the index
which(nstart == max(nstart))
```

```
## Massachusetts Ave & Dupont Circle NW
##                                  227
```

Most trips originated from Massachusetts Ave & Dupont Circle NW.

ii.

```r
nend <- table(bikes$End.Station)
which(nend == max(nend))
```

```
## Massachusetts Ave & Dupont Circle NW
##                                  227
```

It is the same station Massachusetts Ave & Dupont Circle NW.

(d)

```r
# count the number of trips that bike was not returned
sum(bikes$End.Station == "")
```

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

```r
# check the duration of those trips
bikes$Duration[bikes$End.Station == ""]
```

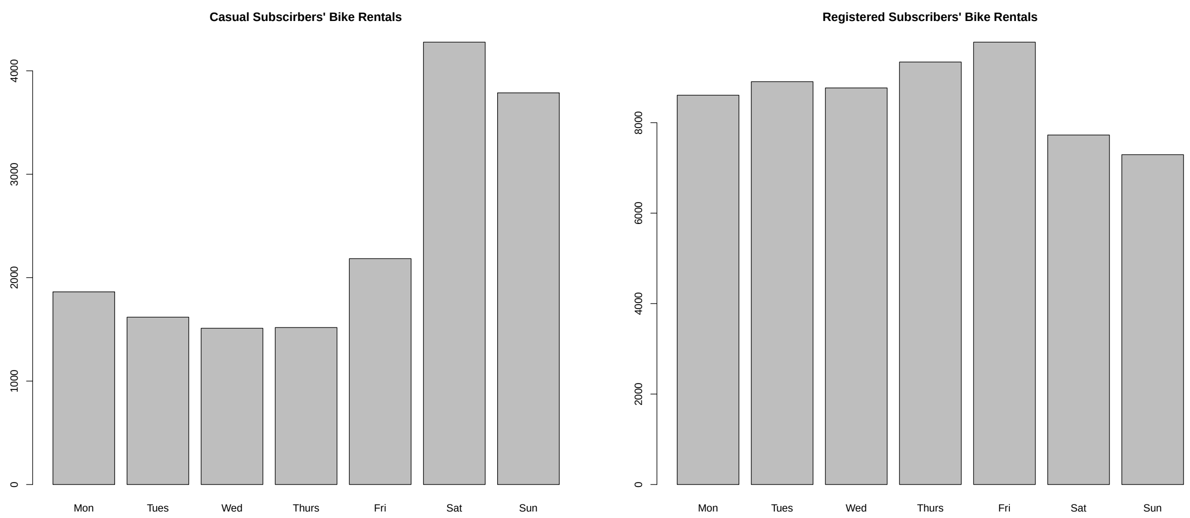
```
## [1] 120
```

```r
# set the duration to NA
bikes$Duration[bikes$End.Station == ""] <- NA
```

Among all 77186 trips only 1 bike rented was not returned, so it is very rare that bike do not get returned. The duration for that trip is 120 seconds.

(e)

```r
wdorder <- c("Mon", "Tues", "Wed", "Thurs", "Fri", "Sat", "Sun")
par(mfrow = c(1, 2))
# select the casual subscribers and plot the bar plot
barplot(table(bikes$wday[bikes$Subscriber.Type == "Casual"])[wdorder],
    main = "Casual Subscirbers' Bike Rentals")
# select the registered subscribers and plot the bar plot
barplot(table(bikes$wday[bikes$Subscriber.Type == "Registered"])[wdorder],
    main = "Registered Subscribers' Bike Rentals")
```



The casual subscrubers tend to rent the bikes on weekends, and the registered subscribers tend to rent the bikes every day of the week, and are more likely to rent bikes on weekdays.

Problem 3

(a) read the data.

```
titanic <- read.table(file = "http://maitra.public.iastate.edu/stat579/datasets/titanic.txt",
    header = T, sep = ",")
```

(b) cross-classify the passengers.

```
# cross-classify the gender and passenger class
table(titanic$Sex, titanic$PClass)
```

```
##
##          1st 2nd 3rd
##   female 143 107 212
##   male   179 173 499
```

```
# cross-classify the gender and passenger class stratified by survival status
table(titanic$Sex, titanic$PClass, titanic$Survived)
```

```
## , ,  = 0
##
##
##          1st 2nd 3rd
##   female   9  13 132
##   male   120 148 441
##
## , ,  = 1
##
##
##          1st 2nd 3rd
##   female 134  94  80
##   male    59  25  58
```

We can see there were more male passengers, especially in the 3rd class. most female passengers in the 1st and 2nd class survived. Most male passengers did not survive. Female passengers in 3rd class have higher survival rate than that of the male passengers in the 3rd class, but is much lower than that of the female passengers in 1st and 2nd class. The overall survival rate is much lower for the 3rd class passengers.

(b)

```
F0 <- titanic$Age[titanic$Sex == "female" & titanic$Survived == 0]
F1 <- titanic$Age[titanic$Sex == "female" & titanic$Survived == 1]
M0 <- titanic$Age[titanic$Sex == "male" & titanic$Survived == 0]
M1 <- titanic$Age[titanic$Sex == "male" & titanic$Survived == 1]

# mean age of female passengers that survived with NA value removed
muF0 <- mean(F0, na.rm = T)
muF0
```

```
## [1] 24.90141
```

```r
# mean age of female passengers that did not survive with NA value removed
muF1 <- mean(F1, na.rm = T)
muF1
```

```
## [1] 30.86714
```

```r
# difference of mean age between female passengers that survived and that did not
difff <- muF0 -muF1
difff
```

```
## [1] -5.965734
```

```r
# number of female passengers that survived and whose age is avalaible
fn0 <- sum(!is.na(F0))
# number of female passengers that did not survive and whose age is avalaible
fn1 <- sum(!is.na(F1))

# standard error of mean age of female passengers that survived with NA value removed
SEmuF0 <- sd(F0, na.rm = T) / sqrt(fn0)
SEmuF0
```

```
## [1] 1.548272
```

```r
# standard error of mean age of female passengers that did not survive with NA value removed
SEmuF1 <- sd(F1, na.rm = T) / sqrt(fn1)
SEmuF1
```

```
## [1] 1.01999
```

```r
# mean age of female passengers that survived with NA value removed
muM0 <- mean(M0, na.rm = T)
muM0
```

```
## [1] 32.32078
```

```r
# mean age of male passengers that did not survive with NA value removed
muM1 <- mean(M1, na.rm = T)
muM1
```

```
## [1] 25.95188
```

```r
# difference of mean age between male passengers that survived and that did not
diffm <- muM0 - muM1
diffm
```

```
## [1] 6.368905
```

```r
# number of male passengers that survived and whose age is avalaible
mn0 <- sum(!is.na(F0))
# number of male passengers that did not survive and whose age is avalaible
mn1 <- sum(!is.na(F1))

# standard error of mean age of female passengers that survived with NA value removed
SEmuM0 <- sd(M0, na.rm = T) / sqrt(mn0)
SEmuM0
```

```
## [1] 1.566366
```

```r
# standard error of mean age of female passengers that did not survive with NA value removed
SEmuM1 <- sd(M1, na.rm = T) / sqrt(mn1)
SEmuM1
```

```
## [1] 1.050158
```

We use t-test to check if there is significant difference in mean age between the female passengers that survived and that did not. We need to make an assumption that they are normally distributed, and ages are all independent. Denote the mean age of female passsengers that survived μ_{F_1}, those that did not survive μ_{F_0}, then

$H_0 : \mu_{F_1} = \mu_{F_0}$

$H_a : \mu_{F_1} \neq \mu_{F_0}$

```r
t.test(F0[!is.na(F0)], F1[!is.na(F1)])
```

```
##
##  Welch Two Sample t-test
##
## data:  F0[!is.na(F0)] and F1[!is.na(F1)]
## t = -3.2177, df = 135.67, p-value = 0.001617
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.632324 -2.299144
## sample estimates:
## mean of x mean of y
##  24.90141  30.86714
```

The p-value is very small, thus we decide to reject the mull hypothesis and conclude that there is difference in the mean age between the female passengers that did not survive and that did survive.

We use t-test to check if there is significant difference in mean age between the male passengers that survived and that did not. We need to make an assumption that they are normally distributed, and ages are all independent. Denote the mean age of male passsengers that survived μ_{M_1}, those that did not survive μ_{M_0}, then

$H_0 : \mu_{M_1} = \mu_{M_0}$

$H_a : \mu_{M_1} \neq \mu_{M_0}$

```r
t.test(M0[!is.na(M0)], M1[!is.na(M1)])
```

```
##
##  Welch Two Sample t-test
##
## data:  M0[!is.na(M0)] and M1[!is.na(M1)]
## t = 3.7011, df = 132.84, p-value = 0.0003134
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.965201 9.772608
## sample estimates:
## mean of x mean of y
##  32.32078  25.95188
```

The p-value is very small, thus we decide to reject the mull hypothesis and conclude that there is difference in the mean age between the male passengers that did not survive and that did survive.