36-711 – Homework 1 Solutions

1.
```
>depr <- read.table("C:\\temp\\711\\depression.dat", header=T)

> names(depr)
[1] "Hospt"   "Treat"   "Outcome" "Time"    "AcuteT"  "Age"
"Gender"

> dimnames(depr)
[[1]]:
  [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"
 [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"
 [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"
 [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"
 [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
 [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"
 [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"
 [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"
 [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
[109] "109"

[[2]]: [1] "Hospt"   "Treat"   "Outcome" "Time"    "AcuteT"  "Age"
"Gender"

> dim(depr)
[1] 109   7
```

2.
```
> summary(depr)
      Hospt           Treat            Outcome            Time
    Min.:1.000      Min.:0.0000      Min.:0.0000      Min.:  0.000
 1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  6.286
  Median:5.000    Median:1.0000    Median:1.0000    Median: 31.860
    Mean:3.927      Mean:0.9725      Mean:0.5229      Mean: 46.520
 3rd Qu.:5.000   3rd Qu.:2.0000   3rd Qu.:1.0000   3rd Qu.: 78.430
    Max.:6.000      Max.:2.0000      Max.:1.0000      Max.:206.000
      AcuteT          Age             Gender
    Min.: 47.0      Min.:20.00      Min.:1.000
 1st Qu.:133.0   1st Qu.:29.00   1st Qu.:1.000
  Median:165.0    Median:39.00    Median:1.000
    Mean:185.7      Mean:40.26      Mean:1.349
 3rd Qu.:214.0   3rd Qu.:51.00   3rd Qu.:2.000
    Max.:512.0      Max.:65.00      Max.:2.000

> sqrt(apply(depr, 2, var))
    Hospt    Treat   Outcome     Time   AcuteT      Age    Gender
 1.879255 0.810335 0.5017807 45.14164 88.61862 12.86209 0.4787357
```

apply(depr, 2, var) applies variance operation on columns of depr. Or alternatively, you can use

```
for(i in names(depr)) {
    print(summary(depr[,i]))
    print(sqrt(var(depr[,i])))
}
```

3. (a) Other variables are categorical, their values are nominal. They can be treated as factors.
   (b) table() generates the same results as below:

```
> summary(as.factor(depr$Hospt))
  1  2  3  5  6
 18 17 10 38 26
> summary(as.factor(depr$Treat))
  0  1  2
 37 38 34
> summary(as.factor(depr$Outcome))
  0  1
 52 57
> summary(as.factor(depr$Gender))
  1  2
 71 38
```

   We observe that more observations are from hospital 5 and 6 than other 3 hospitals; observations are evenly distributed among treatment groups and outcome groups; there are roughly twice amount of female patients as much as male patient.

4. ```
   > attach(depr)
   > search()
   ```

   After depr is attached as a database, variables are easier to access. To access variable "Hospt", you can use "Hospt" instead of "depr$Hospt".

5. ```
   > stem(Time)
   N = 109   Median = 31.857 Quartiles = 6.286, 78.429

   Decimal point is 1 place to the right of the colon

       0 : z0001111111222333333455555666778899
       1 : 124666778
       2 : 01226778
       3 : 01236688
       4 : 0024566
       5 : 016
       6 : 4778
       7 : 2455888889
       8 : 36
       9 : 368
      10 : 02235555667
      11 : 025
      12 : 57
   ```

```
   13 : 1
   14 :
   15 : 5
   16 : 5
   17 :
   18 :
   19 :
   20 : 6

> stem(AcuteT)
N = 109   Median = 165 Quartiles = 133, 214

Decimal point is 2 places to the right of the colon

   0 : 555677899999
   1 : 0000111233333334444444444
   1 : 5555555556666666677777788888889999
   2 : 000001111111233444
   2 : 556678899
   3 : 0124
   3 : 57

High:  399  428  441  476  512

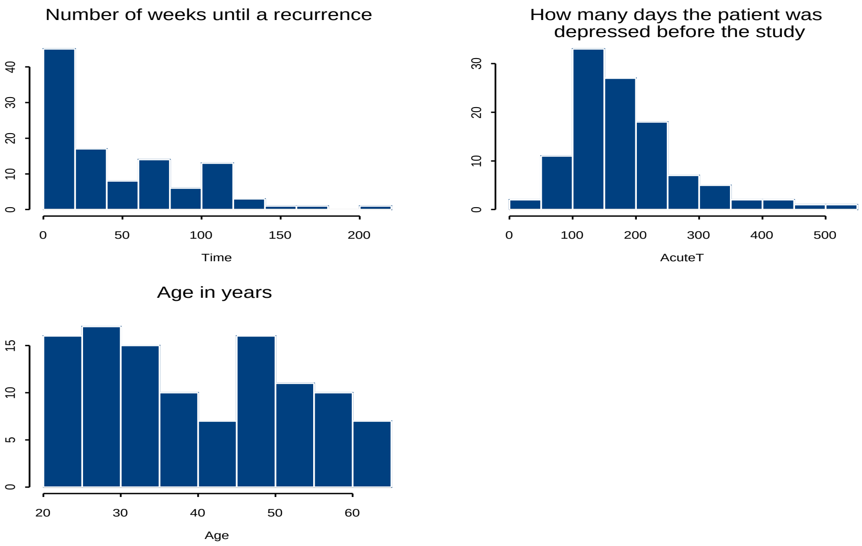
> stem(Age)
N = 109   Median = 39 Quartiles = 29, 51

Decimal point is 1 place to the right of the colon

   2 : 022222333333334
   2 : 56677788889999
   3 : 00001111223333344
   3 : 556678899
   4 : 00011334
   4 : 55666677889999
   5 : 0000111222344
   5 : 556777788
   6 : 00011123
   6 : 55
```

6.
```
> par(mfrow = c(2, 2))
> hist(Time)
> title("Number of weeks until a recurrence")
> hist(AcuteT)
> title("How many days the patient was\n depressed before the study")
> hist(Age)
> title("Age in years")
```

See the following figure.



7.
```
> table(Hospt, Treat)
    0  1  2
1   7  7  4
2   5  7  5
3   4  2  4
5  12 13 13
6   9  9  8
```

Treatments are NOT evenly distributed across hospitals, hospital 5 (and 6) has much more observations than others. However, within each hospital, numbers of each treatments are roughly equal.

ratios of outcomes. apply(bozo[, , i], 1, sum) is to sum each row (each Hospt) over Outcome for the ith Treat. bozo[, , i]/apply(bozo[, , i], 1, sum) gives percentage of good and bad Outcomes within the same Hospt and Treat. round(bozo,2) rounds all entries in bozo to 2 siginifcant digits.

```
> bozo <- table(Hospt, Outcome, Treat)
> for(i in 1:3) bozo[,  , i] <- bozo[,  , i]/apply(bozo[,  , i], 1, sum)
> round(bozo, 2)


, , 0
     0    1
1 0.57 0.43
2 0.40 0.60
3 0.25 0.75
5 0.50 0.50
6 0.11 0.89
```

```
, , 1
     0    1
1 0.86 0.14
2 0.57 0.43
3 1.00 0.00
5 0.85 0.15
6 0.44 0.56


, , 2
     0    1
1 0.25 0.75
2 0.40 0.60
3 0.00 1.00
5 0.54 0.46
6 0.12 0.88
```

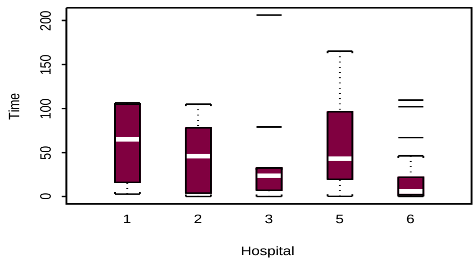
Treatment 1 works better than 0 and 2 in all hospitals. In most hospitals, treatment 0 is about the same as treatment 2 (placebo).

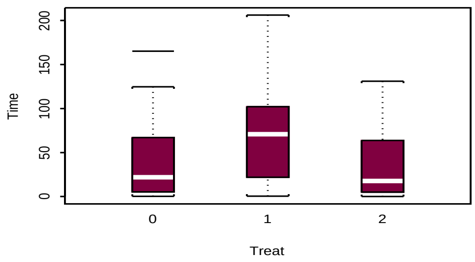
8.
```
> par(mfrow = c(1, 1))
> plot(as.factor(Hospt), Time, xlab = "Hospital", ylab = "Time")
> boxplot(split(Time, Hospt), xlab = "Hostpital", ylab = "Time")
```

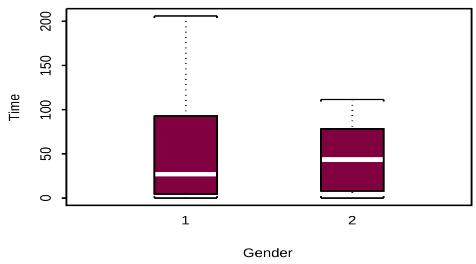
These two methods produce the same parallel box plot.

The time of recurrence are not much different (mean and spread) among hospital 1, 2, and 5, while hospital 3 and 6 have much smaller mean and variance.

We can use this method to check relationship between Treat and Time, and Gender and Time. It again shows that Treat 1 also prolonged the recurrence time.

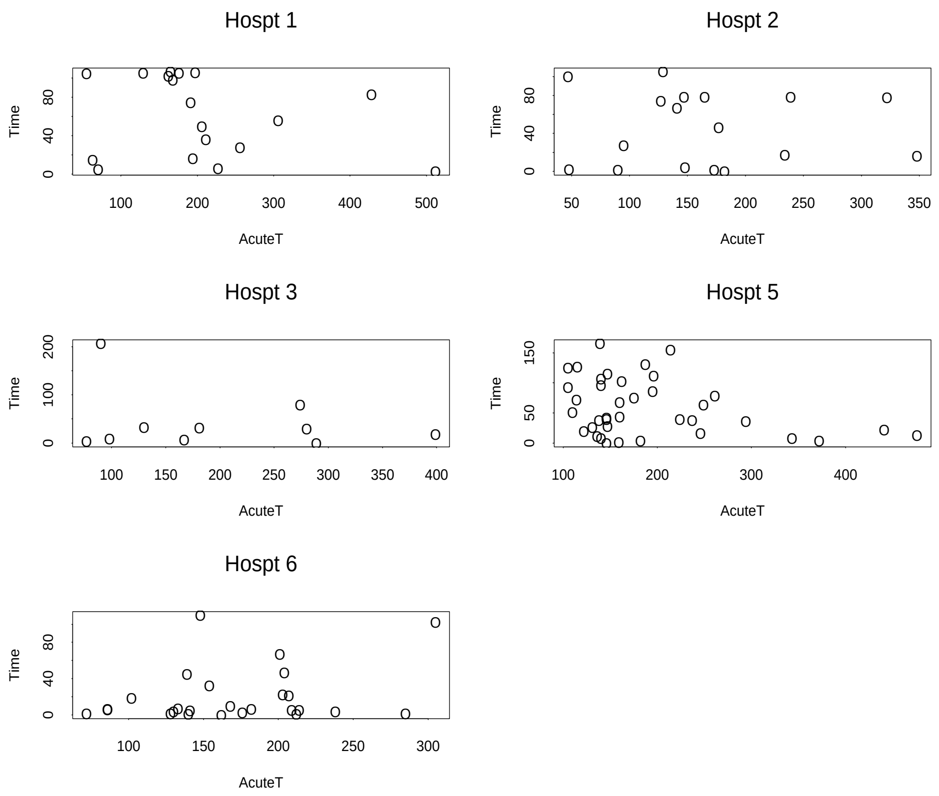






9. ```
   > motif("-g 400x400")
   > par(mfrow = c(3,2))
   > for (i in unique(Hospt)) plot(AcuteT[Hospt==i],Time[Hospt==i],
   + main=paste("Hospt",i),xlab="AcuteT",ylab="Time",pch="O")
   > dev.print(postscript,command="cat > hw1-3.eps",onefile=F)
   ```

   they show no apparent patterns. Only in hosptial 5, those patients with very long AcuteT tend to have very short recurrence time.



This method can also be used to check relationships between Age and recurrence Time.
   ```