STAT 40001/MA 59800 Statistical Computing/ Computational Statistics Fall 2013
Homework 4- Solution

Name:

Due : October 3, 2013 PUID:

Instruction: Please submit your R code along with a brief write-up of the solutions. Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) Data from a study comparing brain size and intelligence is available on the DASL web site *http://lib.stat.cmu.edu/DASL/Datafiles/Brainsize.html.*
a) Import the Data set in R-readable format(You may first save and then import it using `read.table`)
b) Print first 5 observations.
c) It appears that there are few missing values. Identify the missing values.
d) Calculate the summary of each of the variables.

Solution: We have saved the data in the C: directory and give the name brain We imported the data as below:

a)

```
data=read.table("C://Desktop//STAT4001//brain.txt", header=T, na.strings="")
```

b)

```
>  head(data,5)
  Gender FSIQ VIQ PIQ Weight Height MRI_Count
1 Female  133 132 124    118   64.5    816932
2   Male  140 150 124     NA   72.5   1001121
3   Male  139 123 150    143   73.3   1038437
4   Male  133 129 128    172   68.8    965353
5 Female  137 132 134    147   65.0    951545
```

c)

```
 > which (is.na(data), arr.ind = T)
     row col
[1,]   2   5
[2,]  21   5
[3,]  21   6
```

d) > summary(data)

```
    Gender        FSIQ              VIQ              PIQ              Weight
 Female:20   Min.   : 77.00   Min.   : 71.0   Min.   : 72.00   Min.   :106.0
 Male  :20   1st Qu.: 89.75   1st Qu.: 90.0   1st Qu.: 88.25   1st Qu.:135.2
             Median :116.50   Median :113.0   Median :115.00   Median :146.5
             Mean   :113.45   Mean   :112.3   Mean   :111.03   Mean   :151.1
             3rd Qu.:135.50   3rd Qu.:129.8   3rd Qu.:128.00   3rd Qu.:172.0
             Max.   :144.00   Max.   :150.0   Max.   :150.00   Max.   :192.0
                                                               NA's   :2
     Height         MRI_Count
 Min.   :62.00   Min.   : 790619
 1st Qu.:66.00   1st Qu.: 855919
 Median :68.00   Median : 905399
 Mean   :68.53   Mean   : 908755
 3rd Qu.:70.50   3rd Qu.: 950078
 Max.   :77.00   Max.   :1079549
 NA's   :1
```

Q.N. 2) Access the data from url *http://www.stat.berkeley.edu/users/statlabs/data/babies.data* and store the information in an object named **BABIES** using the function *read.table()*
a) Create a CLEAN data set that removes subjects if any observations on the subject are "unknown". Note that if the values are unknown **bwt, gestation, parity, height, weight** and **smoke** are quoted as 999, 999, 9, 99, 999, and 9 respectively. Store the modified data set in an object named CLEAN.
b) Create side-by-side boxplots to compare the birth weights of babies for both smoking and non-smoking mothers.
c) Calculate the five number summaries of the birth weights of babies of both smoking and non-smoking mothers.

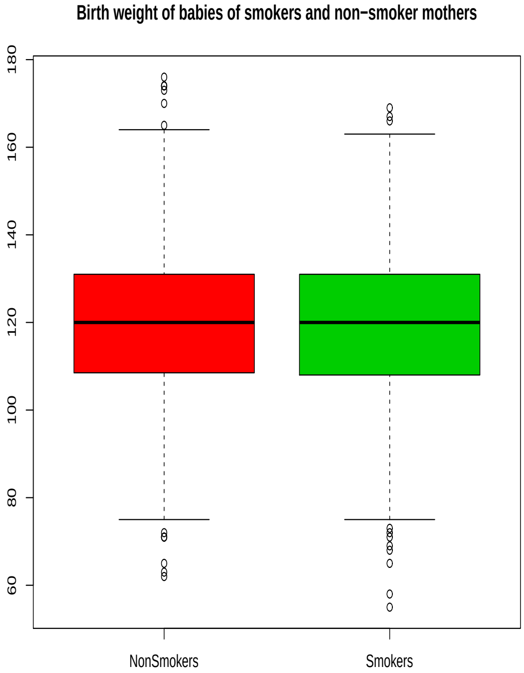
Solution: Solution: The R code below will be used to answer these questions

```
> site<-"http://www.stat.berkeley.edu/users/statlabs/data/babies.data"
> BABIES<-read.table(site, header=T)
> attach(BABIES)
> CLEAN<-subset(BABIES, bwt!=999 & gestation!=999 & parity!=9 & height!=99 & weight!=999 & smoke!=9)
> dim(BABIES)
[1] 1236    7
> dim(CLEAN)
[1] 1175    7
```

It appears that the original data has 1236 observations and the CLEAN data has 1175 observations. In fact we could have deleted one more observation with age=99 but the question doesn't mention this.

b)

```
> Smokers=subset(CLEAN$bwt,smoke=="1")
> NonSmokers=subset(CLEAN$bwt,smoke=="0")
> boxplot(NonSmokers,Smokers,col=c(2,3),main="Birth weight of babies of smokers and non-sm
 names=c("NonSmokers","Smokers"))
```



c) The five number summary for the weight of new born babies with smoker and nonsmoker mothers is given as

```
> fivenum(Smokers)
[1]  55 108 120 131 169
> fivenum(NonSmokers)
[1]  62.0 108.5 120.0 131.0 176.0
```

Q.N. 3) Suppose that the population of adult male bears has weights that are approximately normally distributed with average 350 lbs and standard deviation of 75 lbs. What is the probability that a randomly observed male bear weighs more than 450 lbs?

Solution: Since the weights of adult male bears are approximately normally distributed we can use the R code below to find the probability of a randomly observed male bear weight greater than 450

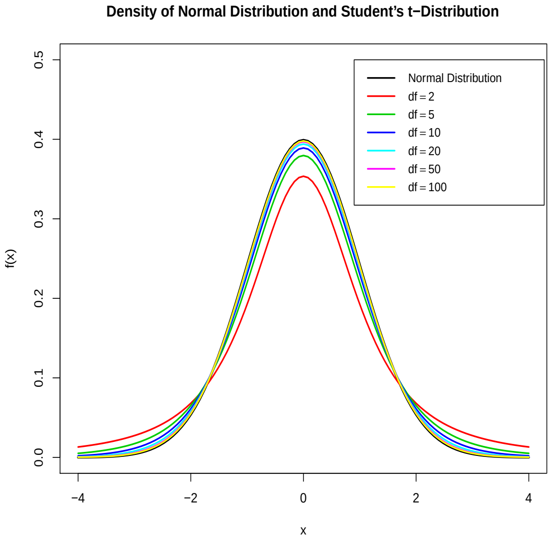
```
> 1-pnorm(450,350,75)
[1] 0.09121122
```

Q.N. 4) If $\boldsymbol{X \sim \chi^2_{10}}$.
a) Calculate $\boldsymbol{P(X \leq 8)}$
b) Calculate $\boldsymbol{P(X > 6)}$
c) Calculate $\boldsymbol{a}$ so that $\boldsymbol{P(X < a) = 0.05}$**.**
Solution:

```
> pchisq(8,10)
[1] 0.3711631
> 1-pchisq(6,10)
[1] 0.8152632
> qchisq(0.05,10)
[1] 3.940299
```

Q.N. 5) For the t-distribution we can see that as the degrees of freedom get large the density approaches the normal. To investigate , plot the standard normal density and add densities of the t-distributions with different degrees of freedom.
Solution: We can use R code below to demonstrate that t distribution converges to standard normal as the degrees of freedom increases.

```
> curve(dnorm(x),-4, 4, col=1, lwd=3,ylim=c(0,0.5),ylab="f(x)",xlab="x",
 main="Density of Normal Distribution and Student’s t-Distribution")
> curve(dt(x,2),-4,4,col=2, lwd=2,add=T)
> curve(dt(x,5),-4, 4, col=3,lwd=2, add=T)
> curve(dt(x,10),-4, 4, col=4, add=T,lwd=2)
> curve(dt(x,20),-4, 4, col=5, add=T, lwd=2)
> curve(dt(x,50),-4, 4, col=6, add=T, lwd=2)
> curve(dt(x,100),-4, 4, col=7, add=T, lwd=2)
> legend(0.9, 0.5, lwd=2,legend=c(expression("Normal Distribution"),expression(df==2),
expression(df==5),expression(df==10), expression(df==20),expression(df==50),
expression(df==100)),cex=0.9, col=c(1,2,3,4,5,6,7))
```


Density of Normal Distribution and Student’s t–Distribution
Normal Distribution
df = 2
df = 5
df = 10
df = 20
df = 50
df = 100
f(x)
x