**STAT 40001/MA59800 Statistical Computing Fall 2017**

**Lab-10**

1. The data are from a national sample of 6000 households with a male head earning less than $15,000 annually in 1966. The data were classified into 39 demographic groups for analysis. The study was undertaken in the context of proposals for a guaranteed annual wage (negative income tax). At issue was the response of labor supply (average hours) to increasing hourly wages. The study was undertaken to estimate this response from available data.

Data are available on the DASL web site

<http://lib.stat.cmu.edu/DASL/Datafiles/wagesdat.html>

Import the Data set in R-readable format (You may first save and then import it using read.table). It appears that missing values are indicated by “\*”.

1. Create a clean dataset by replacing “\*” by NA.

> data <- read.table("C:\\Users\\Administrator\\Desktop\\statistical computing\\lab10\\wagesdat.txt",header = T,na.strings = "\*")

> head(data)

HRS RATE ERSP ERNO NEIN ASSET AGE DEP RACE SCHOOL

1 2157 2.905 1121 291 380 7250 38.5 2.340 32.1 10.5

2 2174 2.970 1128 301 398 7744 39.3 2.335 31.2 10.5

3 2062 2.350 1214 326 185 3068 40.1 2.851 NA 8.9

4 2111 2.511 1203 49 117 1632 22.4 1.159 27.5 11.5

5 2134 2.791 1013 594 730 12710 57.7 1.229 32.5 8.8

6 2185 3.040 1135 287 382 7706 38.6 2.602 31.4 10.7

1. Test the hypothesis that the Average yearly earnings of spouse (ERSP) is greater than $1100.

> t.test(data$ERSP,alternative="greater", mu=1100)

One Sample t-test

data: data$ERSP

t = 0.037937, df = 36, p-value = 0.485

alternative hypothesis: true mean is greater than 1100

95 percent confidence interval:

1025.929 Inf

sample estimates:

mean of x

1101.703

1. Test the hypothesis that the average age of respondent is less than 40 years.

> t.test(data$AGE,alternative="less", mu=40)

One Sample t-test

data: data$AGE

t = -0.9595, df = 38, p-value = 0.1717

alternative hypothesis: true mean is less than 40

95 percent confidence interval:

-Inf 40.49115

sample estimates:

mean of x

39.35128

1. ***faraway*** package in R contains a data set ***prostate*** which describes 97 men with prostate cancer who were due to receive a radical prostatectomy. Test the hypothesis whether the participants are younger than 65 years.

> install.packages("faraway")

> library(faraway)

> t.test(prostate,alternative="less", mu=65)

One Sample t-test

data: prostate

t = -71.419, df = 872, p-value < 2.2e-16

alternative hypothesis: true mean is less than 65

95 percent confidence interval:

-Inf 12.63779

sample estimates:

mean of x

11.40206

1. Information about several colleges and Universities in the United States are provided in the fourth file in the link below

<http://mathforum.org/workshops/sum96/data.collections/datalibrary/data.set6.html>

1. Import the data in R  
   data <- read.csv("C:\\Users\\wu1114\\Desktop\\college.csv")
2. How many college/universities are in each state?  
   > table(data$State)

AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA

4 25 17 5 70 16 19 9 5 30 36 5 29 6 49 42 20 24 20

MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR

56 23 14 36 25 35 15 9 46 8 16 13 27 9 3 101 52 20 18

PA RI SC SD TN TX UT VA VT WA WI WV WY

83 8 26 11 34 60 6 40 14 16 29 17 1

1. Create a clean data by deleting the missing values.  
   data <- na.omit(data)  
   > dim(data)

[1] 206 23

1. In the clean data set how many are private and how many public University/colleges are included?  
   > table(data[,3])

1 2

46 160

1. Test the hypothesis the additional fees (other than tuition fee) charged by public colleges/universities is more than $400.  
   > t.test(data[data[,3]==1,18],alternative="greater",mu=400)

One Sample t-test

data: data[data[, 3] == 1, 18]

t = 1.2514, df = 45, p-value = 0.1086

alternative hypothesis: true mean is greater than 400

95 percent confidence interval:

381.44 Inf

sample estimates:

mean of x

454.2609