**STAT 40001/MA 59800 Statistical Computing Fall 2017**

**Lab-12**

1. The internet traffic data are available in the link below

<http://www.amstat.org/publications/jse/datasets/packetdata.dat.txt>

The description of the variables can be obtained below

<https://ww2.amstat.org/publications/jse/datasets/packetdata.txt>

1. Identify the variables and the dimension of the dataset.

> data = read.table("C:\\Users\\Administrator\\Desktop\\statistical computing\\lab12\\task1.txt",header = T)

> names(data)

[1] "timestamp" "source" "destination" "sourceport" "destport" "databytes"

> dim(data)

[1] 49999 6

1. Print first five observation of the dataset.

> head(data,5)

timestamp source destination sourceport destport databytes

1 0.416754 1 2 1223 2046 0

2 0.418705 2 3 1985 20 0

3 0.420657 4 5 119 3849 5

4 0.426512 3 2 20 1985 512

5 0.427488 3 2 20 1985 512

1. Is there any missing value?

> sum(is.na(data))

[1] 0

1. The website below provides the dataset related to the study of the maternal smoking and infant health.

<http://www.stat.berkeley.edu/~statlabs/data/babiesI.data>

There are two variables

bwt=Birth weight in ounces

smoke=Smoking status of mother 0=not now, 1=yes now, 9=unknown

1. Import the data set in R

> data = read.table("http://www.stat.berkeley.edu/~statlabs/data/babiesI.data",header=TRUE)

1. How many observations have smoking status unknown?

> length(which(smoke==0))

[1] 742

1. CLEAN data set by removing subjects whose smoking status is unknown.

> data1 = data[smoke!=0,]

1. Do we have evidence to prove that the newborn baby will have significantly low weight for a smoker mom than for a non-smoker mom?

> t.test(data[smoke==1,'bwt'],data[smoke==0,'bwt'],alt='less')

Welch Two Sample t-test

data: data[smoke == 1, "bwt"] and data[smoke == 0, "bwt"]

t = -8.5813, df = 1003.2, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -7.222928

sample estimates:

mean of x mean of y

114.1095 123.0472

p-value < 0.05, reject null pothesis.

1. The weight of male and female adult octopuses fished off the coast of Mauritania are provided in the file “Octopus.csv” in the website

<http://www.agrocampus-ouest.fr/igagrocampus-ouest.fr/math/RforStat/>

1. Read the data in R

> data = read.csv("C:\\Users\\Administrator\\Desktop\\statistical computing\\lab12\\Octopus.csv",sep = ';')

> head(data)

Weight Sex

1 300 Female

2 700 Female

3 850 Female

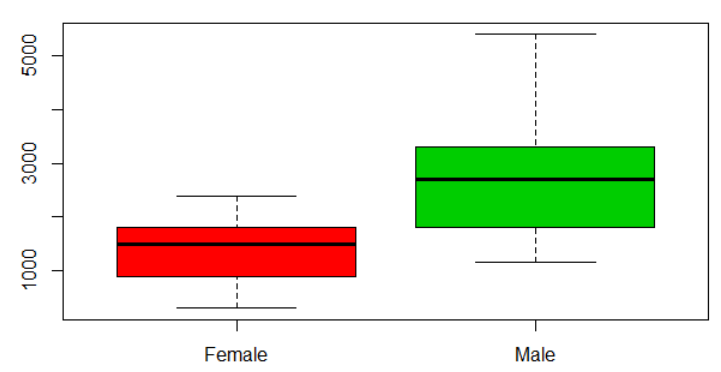
4 900 Female

5 1000 Female

6 1420 Female

1. Compare the two sub-population graphically

> boxplot(data$Weight~data$Sex,col=c(2,3))



1. Calculate the descriptive statistics for each sub-population

> tapply(Weight,Sex,summary)

$F

Min. 1st Qu. Median Mean 3rd Qu. Max.

157.0 160.0 163.0 167.7 173.0 183.0

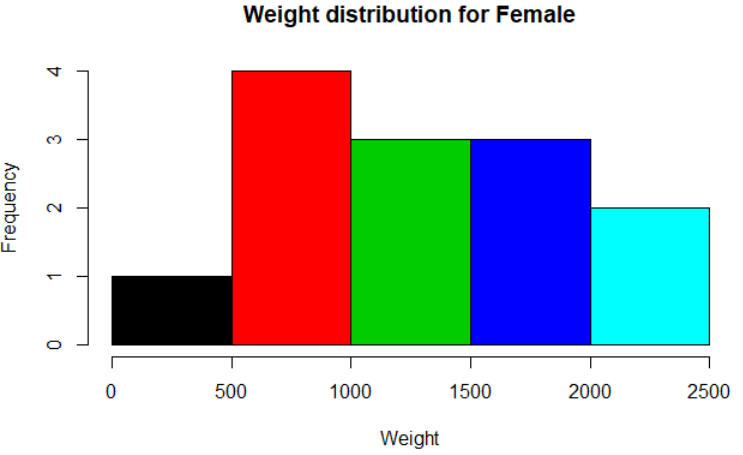
$M

Min. 1st Qu. Median Mean 3rd Qu. Max.

139.0 149.5 154.0 154.2 158.8 170.0

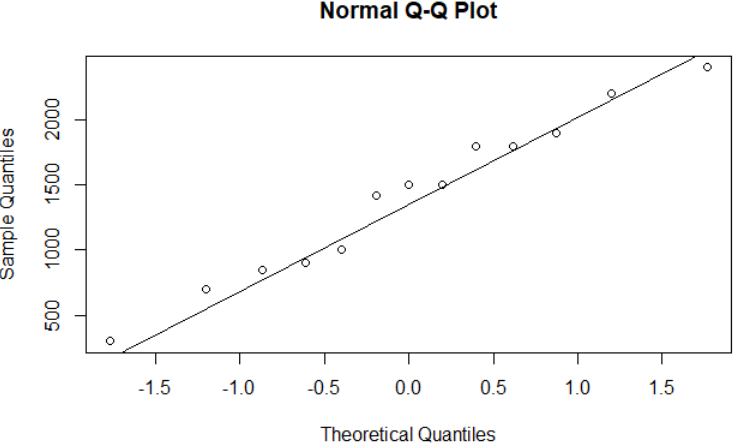
1. Test the normality of the data in each sub-population

> hist(data[data$Sex=='Female','Weight'],c = seq(1:10),main='Weight distribution for Female',xlab = 'Weight')

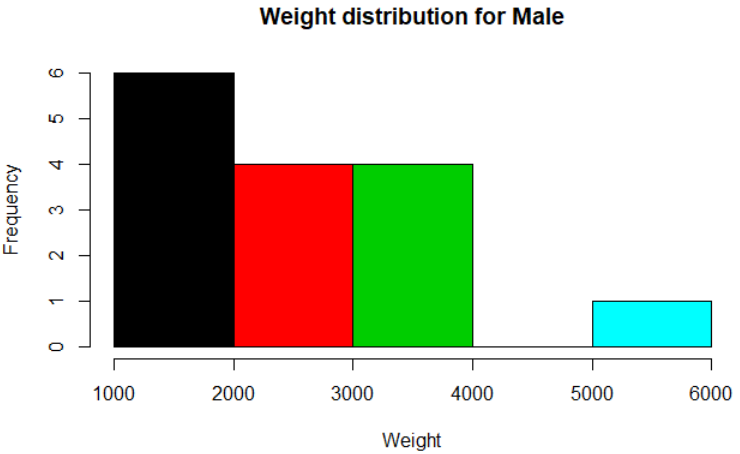


> qqnorm(data[data$Sex=='Female','Weight'])

> qqline(data[data$Sex=='Female','Weight'])

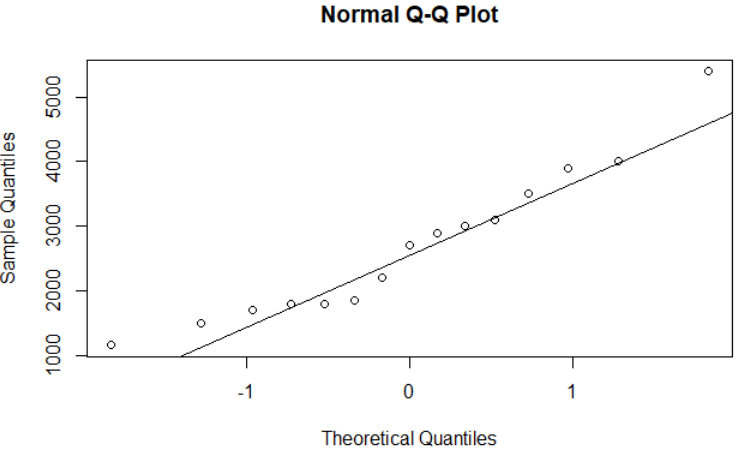


> hist(data[data$Sex=='Male','Weight'],c = seq(1:10),main='Weight distribution for Male',xlab = 'Weight')



> qqnorm(data[data$Sex=='Male','Weight'])

> qqline(data[data$Sex=='Male','Weight'])



1. Test the equality of the variances.

> var.test(data[data[,'Sex']=='Female','Weight'],data[data[,'Sex']=='Male','Weight'])

F test to compare two variances

data: data[data[, "Sex"] == "Female", "Weight"] and data[data[, "Sex"] == "Male", "Weight"]

F = 0.28833, num df = 12, denom df = 14, p-value = 0.03713

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.09452959 0.92444666

sample estimates:

ratio of variances

0.2883299

p-value < 0.05, reject H0

1. Test the equality of means.

> t.test(data[data[,'Sex']=='Female','Weight'],data[data[,'Sex']=='Male','Weight'])

Welch Two Sample t-test

data: data[data[, "Sex"] == "Female", "Weight"] and data[data[, "Sex"] == "Male", "Weight"]

t = -3.7496, df = 22.021, p-value = 0.001107

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2010.624 -578.607

sample estimates:

mean of x mean of y

1405.385 2700.000

p-value < 0.05, reject H0 , not enough evidence to prove that the difference in means is equal to 0

1. The file snore.txt in the link below contains a sample of 100 patients. <http://www.agrocampus-ouest.fr/math/RforStat/snore.txt>
2. Read the data in R

> data = read.csv("C:\\Users\\Administrator\\Desktop\\statistical computing\\lab12\\snore.txt",sep=' ')

> head(data)

age weight size alcohol sex snore tobacco

1 47 71 158 0 M N Y

2 56 58 164 7 M Y N

3 46 116 208 3 M N Y

4 70 96 186 3 M N Y

5 51 91 195 2 M Y Y

6 46 98 188 0 W N N

1. Identify the variables in this study

> names(data)

[1] "age" "weight" "size" "alcohol" "sex" "snore" "tobacco"

1. How many male and how many female have snoring problem?

> attach(data)

> xtabs(~sex+snore)

snore

sex N Y

M 45 30

W 20 5

1. Is there sufficient evidence that tobacco users are more likely to have snoring problem?

> xtabs(~tobacco+snore)

snore

tobacco N Y

N 21 15

Y 44 20

> prop.test(x=c(20,15),n=c(64,36),alt='greater')

2-sample test for equality of proportions with continuity correction

data: c(20, 15) out of c(64, 36)

X-squared = 0.68872, df = 1, p-value = 0.7967

alternative hypothesis: greater

95 percent confidence interval:

-0.2912432 1.0000000

sample estimates:

prop 1 prop 2

0.3125000 0.4166667

p-value > 0.05, do not have enough evidence to reject H0

1. Is there sufficient evidence that more males are tobacco user than female?

> xtabs(~sex+tobacco)

tobacco

sex N Y

M 21 54

W 15 10

> prop.test(x=c(54,10),n=c(75,25),alt='greater')

2-sample test for equality of

proportions with continuity

correction

data: c(54, 10) out of c(75, 25)

X-squared = 7.0023, df = 1, p-value =

0.00407

alternative hypothesis: greater

95 percent confidence interval:

0.1109992 1.0000000

sample estimates:

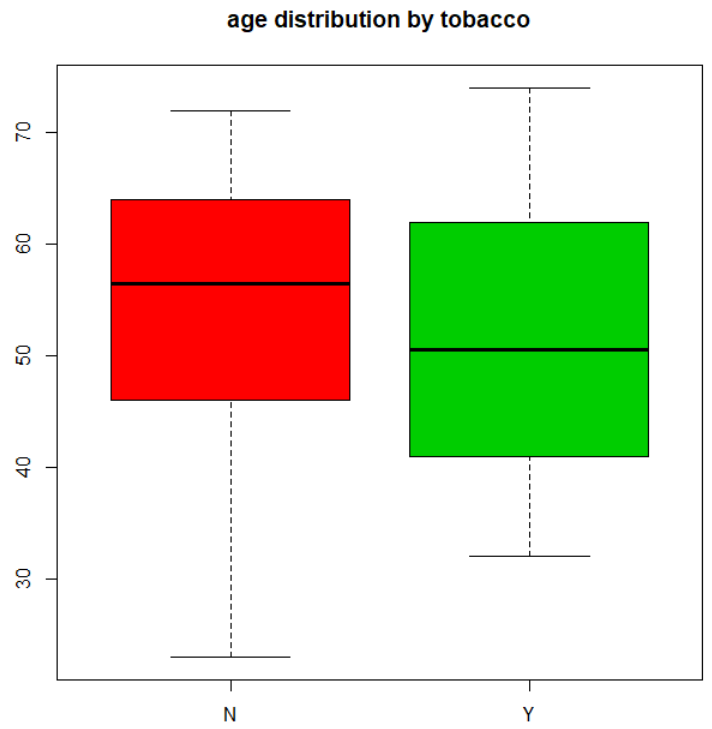
prop 1 prop 2

0.72 0.40

p-value < 0.05, reject the null hypothesis, there is strong evidence that more males are tobacco user than female

1. Display the age distribution by tobacco using status.

> boxplot(data$age~factor(data$tobacco),col=c(2,3),main='age distribution by tobacco')



1. Thirteen Honda Accord are chosen to study the gas mileage. Below is the mpg for these vehicles.

27 26 31 32 30 28 26 24 31 30 23 30 23

1. Test the hypothesis that the variance is not equal to 10.

> data = scan()

1: 27 26 31 32 30 28 26 24 31 30 23 30 23

14:

Read 13 items

> data

[1] 27 26 31 32 30 28 26 24 31 30 23 30 23

> install.packages("TeachingDemos")

> library(TeachingDemos)

> sigma.test(data,sigmasq=10)

One sample Chi-squared test for variance

data: data

X-squared = 12.031, df = 12, p-value = 0.8864

alternative hypothesis: true variance is not equal to 10

95 percent confidence interval:

5.155308 27.319135

sample estimates:

var of data

10.02564

p-value > 0.05, not enough evidence to reject null hypothesis.

1. Construct a 90% confidence interval for the population variance

> sigma.test(data,sigma=10,conf.level=0.9)

One sample Chi-squared test for variance

data: data

X-squared = 1.2031, df = 12, p-value = 7.881e-05

alternative hypothesis: true variance is not equal to 100

90 percent confidence interval:

5.721835 23.020860

sample estimates:

var of data

10.02564

p-value < 0.05, reject the null hypothesis, there is strong evidence that the variance is not equal to 10.

1. Construct a 90% confidence Interval for the population standard deviation.

> zsum.test(mean.x =mean(data), sigma.x =sd(data), n.x = length(data),conf.level = 0.9)$conf.int

[1] 26.32475 29.21371

attr(,"conf.level")

[1] 0.9

1. The following data gives SAT mean scores for math by state for 1989 and 1999 for randomly selected 16 states (Source: The World Almanac and book of facts 2000).

**State 1989 1999**

Arizona 523 525

Connecticut 498 509

Alabama 539 555

Indiana 487 498

Kansas 561 576

Oregon 509 525

Nebraska 560 571

New York 496 502

Virginia 507 499

Washington 515 526

Illinois 539 585

North Carolina 469 493

Georgia 475 482

Nevada 512 517

Ohio 520 568

New Hampshire 510 518

Assuming that the samples comes from a normal distribution,

(a) Test that the mean SAT score for math in 1999 is greater than that in 1989 at. Assume the variances are equal.

> data = read.table("C:\\Users\\Administrator\\Desktop\\statistical computing\\lab12\\task6.txt",sep=',')

> names(data) = c('State',"Year1989","Year1999")

> head(data)

State Year1989 Year1999

1 Arizona 523 525

2 Connecticut 498 509

3 Alabama 539 555

4 Indiana 487 498

5 Kansas 561 576

6 Oregon 509 525

> t.test(data$Year1999,data$Year1989,conf.level = 0.95,var.equal = T)

Two Sample t-test

data: data$Year1999 and data$Year1989

t = 1.3561, df = 30, p-value = 0.1852

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-7.242078 35.867078

sample estimates:

mean of x mean of y

528.0625 513.7500

p-value > 0.05, do not have enough evidence to reject null hypothesis

(b) Test for the equality of the variances at.

> var.test(data$Year1999,data$Year1989,conf.level = 0.95)

F test to compare two variances

data: data$Year1999 and data$Year1989

F = 1.5124, num df = 15, denom df = 15, p-value = 0.4324

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5284078 4.3284913

sample estimates:

ratio of variances

1.512352

p-value > 0.05, do not have enough evidence to reject null hypothesis