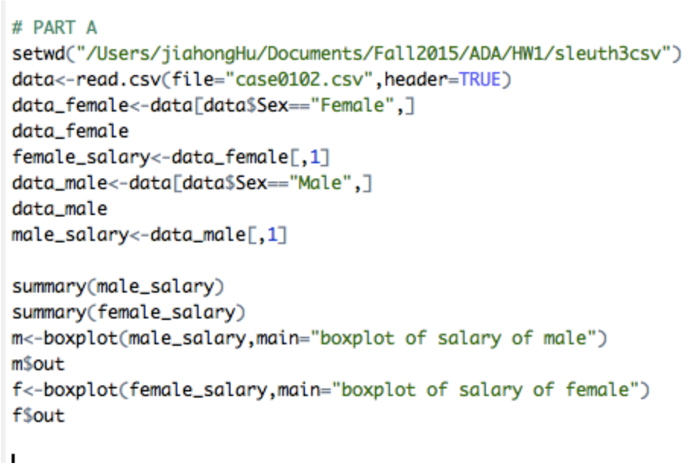
STAT 4201-Advanced Data Analysis
HW1
Name: Jiahong Hu
UNI: jh3561
Fall 2015

Part I

Code:



```
# PART A
setwd("/Users/jiahongHu/Documents/Fall2015/ADA/HW1/sleuth3csv")
data<-read.csv(file="case0102.csv",header=TRUE)
data_female<-data[data$Sex=="Female",]
data_female
female_salary<-data_female[,1]
data_male<-data[data$Sex=="Male",]
data_male
male_salary<-data_male[,1]

summary(male_salary)
summary(female_salary)
m<-boxplot(male_salary,main="boxplot of salary of male")
m$out
f<-boxplot(female_salary,main="boxplot of salary of female")
f$out
```

Definition: An extreme value is considered to be an outlier if it is at least 1.5 IQR below the first Quartile (Q1), or at least 1.5 IQR above the third quartile (Q3).

- The salary of men

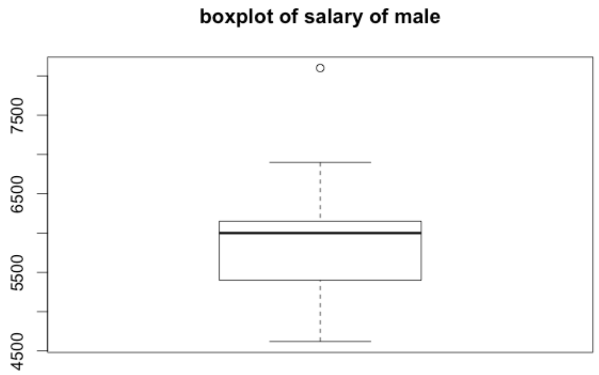
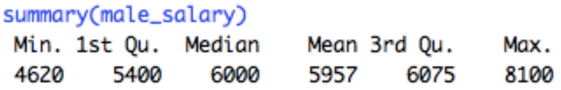


Figure 1.a



```
summary(male_salary)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4620    5400    6000    5957    6075    8100
```

Figure 1.b

The Figure 1.b shows that the average salary of men is $5957, Q1=$5400, Q3=$6075. According to the definition, any observation it is at least 1.5*(Q3-Q1)=$1012.5 below the first Quartile (Q1), or at least $1012.5 above the third quartile (Q3), is considered as an outlier. In this example, $8100 is an outlier since $6075+$1012.5 = $7087.5<$8100. The outlier is indicated by a small hallow circle in the boxplot in Figure 1.a.

- The salary of female

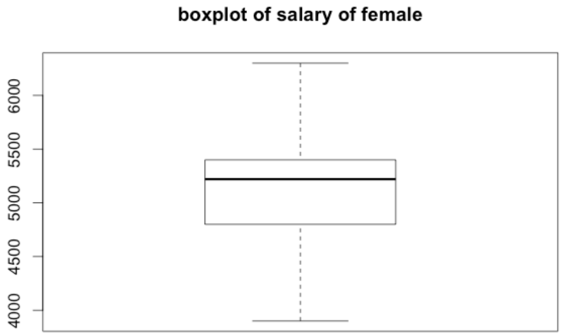
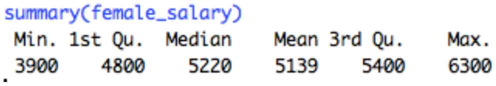


Figure 1.c

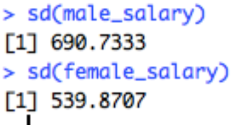


```
summary(female_salary)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3900    4800    5220    5139    5400    6300
```

Figure 1.d

The Figure 1.d shows that the average salary of female is $5139, Q1=$4800, Q3=$5400. According to the definition, any observation it is at least 1.5*(Q3-Q1)=$900 below the first Quartile (Q1), or at least 900 above the third quartile (Q3), is considered as an outlier. In this example, there is no outlier since $4800-$900=$3900 and $5400+$900=$6300. Hence, there is no harrow circle in the boxplot shown in Figure 1.c.

Part II



```
> sd(male_salary)
[1] 690.7333
> sd(female_salary)
[1] 539.8707
```

Figure 2.a

- Sample Mean and Median

In Figure 1.b and 1.d, male salary has sample mean $5957 and sample median $6000; female salary has sample mean $5139 and sample median $5220. The difference between sample means and sample medians for male is larger than that for female. The sample mean and median for female is very close, which indicates a large chance for bell shape distribution without skew.

- Sample Standard deviation

As shown in Figure 2.a, the Standard Deviation for male salary is $690.7333 and is $539.8707 for female salary.

- IQR

As indicated in Figure 1.b, IQR for male salary = 1.5*(Q3-Q1)=1.5*(6075-5400)=$1012.5.

According to Figure 1.d, IQR for female salary = 1.5*(Q3-Q1) = 1.5*(5400-4800)=$900.

Graph:

- Male Salary
    - Histogram

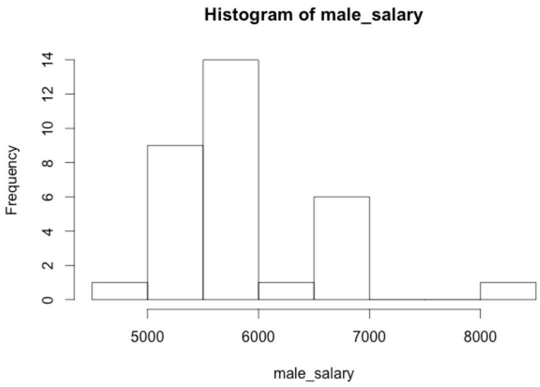


Figure 2.b

- Q-Q Plot

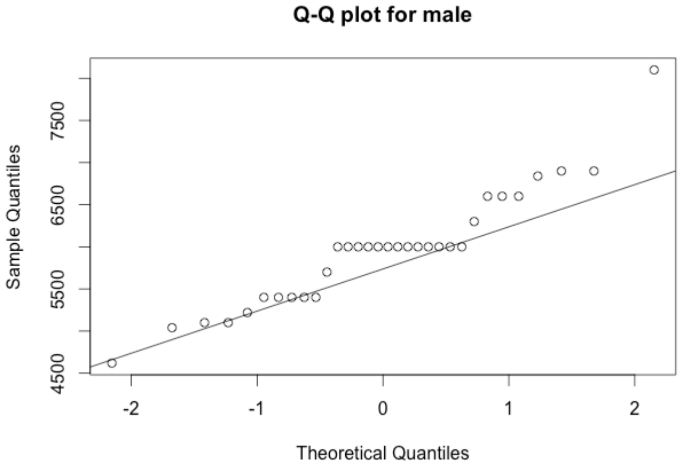


Figure 2.c

As shown in Figure 2.b, the distribution of male salary is not normally distributed but skewed with long right tail. As we know in part I, the observation with salary $8100 is an outlier. As shown in Q-Q plot Figure 2c, when we leave out the outlier, the distribution of male salary does follow closer but not exactly to normal distribution.

- Female Salary
  - Histogram

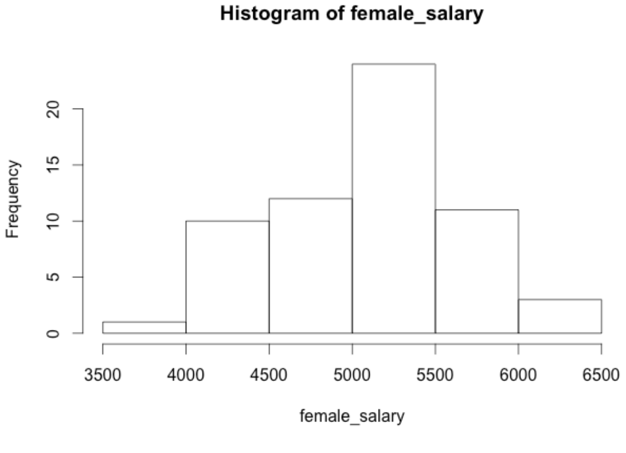


Figure 2.d

- Q-Q Plot

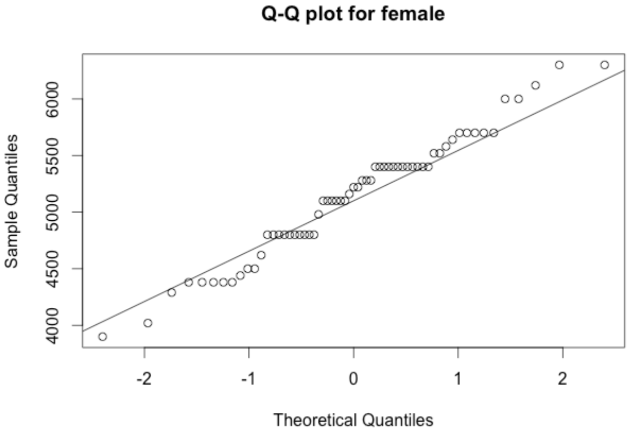


Figure 2.e

As shown in Figure 2.d and 2.e, the female salary has a bell shape and is approximately normally distributed.

- Q-Q plot can also be used to determine if two data sets come from the population with same distribution.

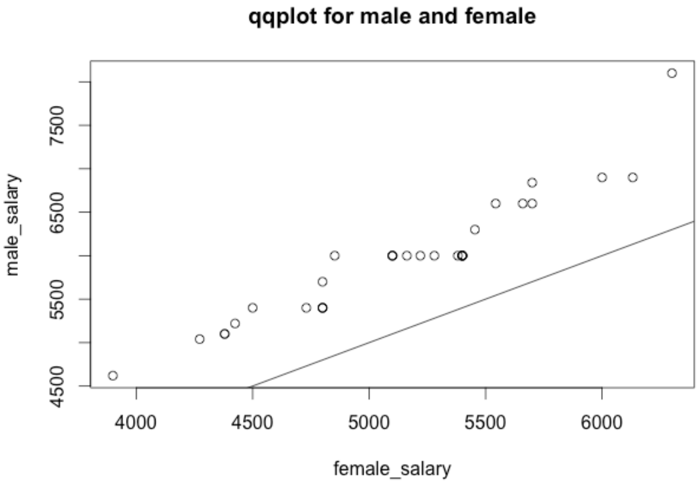


Figure 2.f

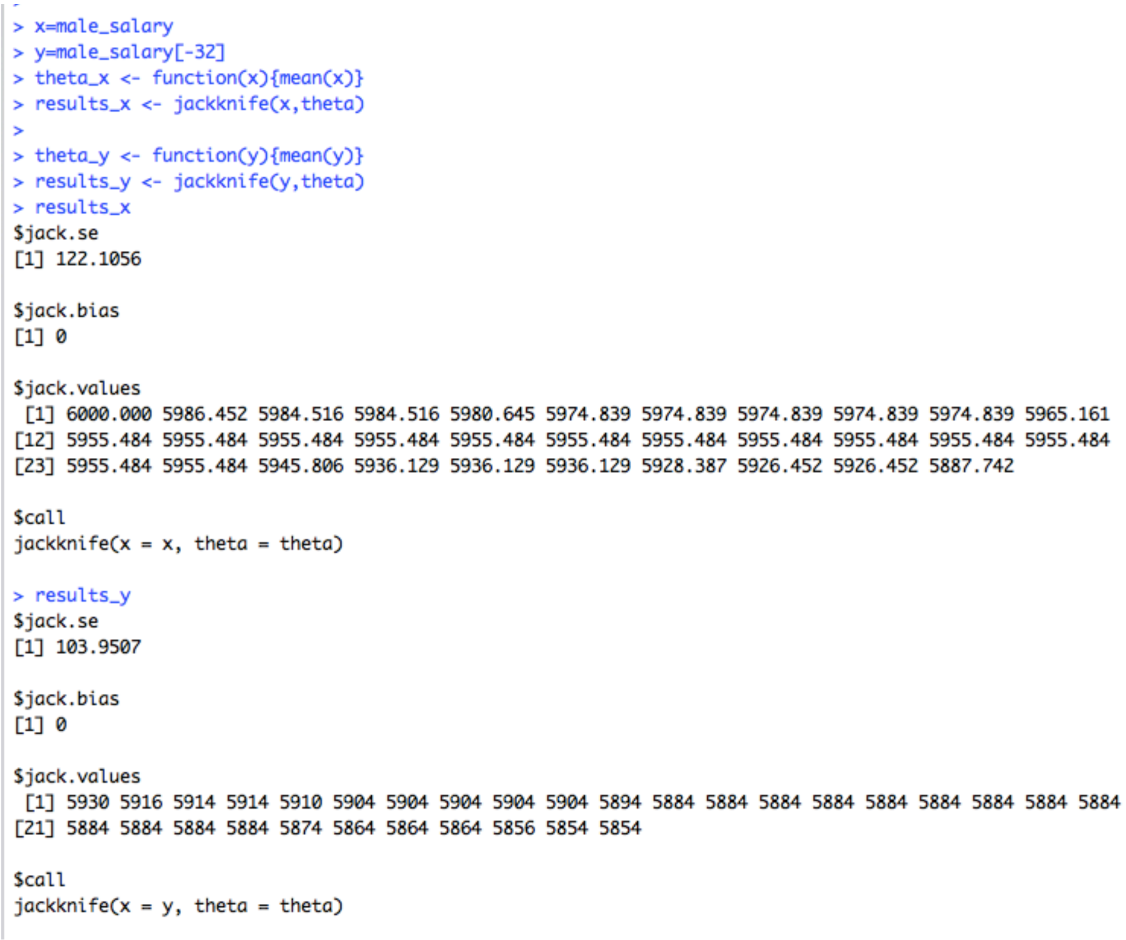
According to Figure 2.f, the Q-Q plot suggests that female salary and male salary have a significant chance of sampling from the two different populations.

Part III

- Jackknife:

    - For Male Salary: The Jackknife method is very sensitive to outlier. I will perform jackknife for each estimator with and without outliers to form a comparison study.

        - Sample Mean

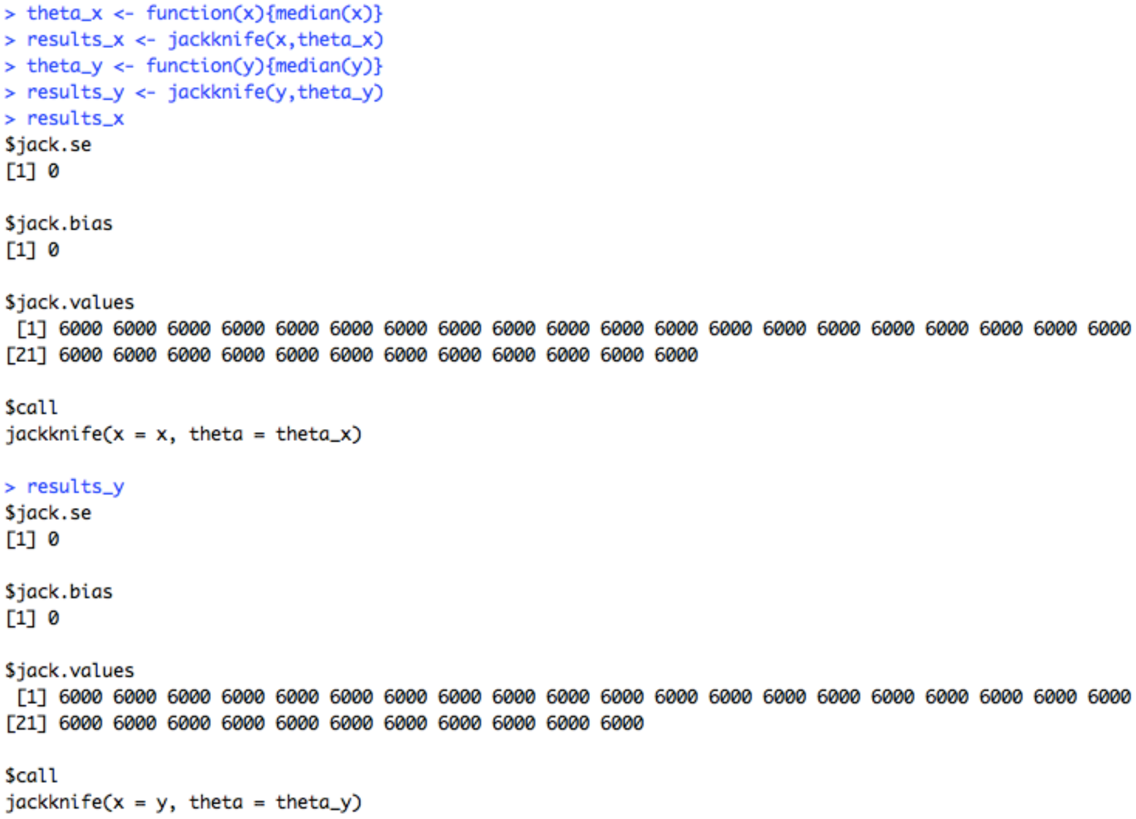


```
> x=male_salary
> y=male_salary[-32]
> theta_x <- function(x){mean(x)}
> results_x <- jackknife(x,theta)
>
> theta_y <- function(y){mean(y)}
> results_y <- jackknife(y,theta)
> results_x
$jack.se
[1] 122.1056

$jack.bias
[1] 0

$jack.values
 [1] 6000.000 5986.452 5984.516 5984.516 5980.645 5974.839 5974.839 5974.839 5974.839 5974.839 5965.161
[12] 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484 5955.484
[23] 5955.484 5955.484 5945.806 5936.129 5936.129 5936.129 5928.387 5926.452 5926.452 5887.742

$call
jackknife(x = x, theta = theta)

> results_y
$jack.se
[1] 103.9507

$jack.bias
[1] 0

$jack.values
 [1] 5930 5916 5914 5914 5910 5904 5904 5904 5904 5904 5894 5884 5884 5884 5884 5884 5884 5884 5884 5884
[21] 5884 5884 5884 5884 5874 5864 5864 5864 5856 5854 5854

$call
jackknife(x = y, theta = theta)
```

Figure 3.a

Figure 3.a shows that the bias under the jackknife method is 0, both with outlier and without outlier. The variance of estimator is (SE)^2=(122.1056)^2=14909.78 with outlier and (103.9507)^2=10805.75 without outlier. The variance is smaller when the outlier is eliminated under the jackknife method.

- Sample Median

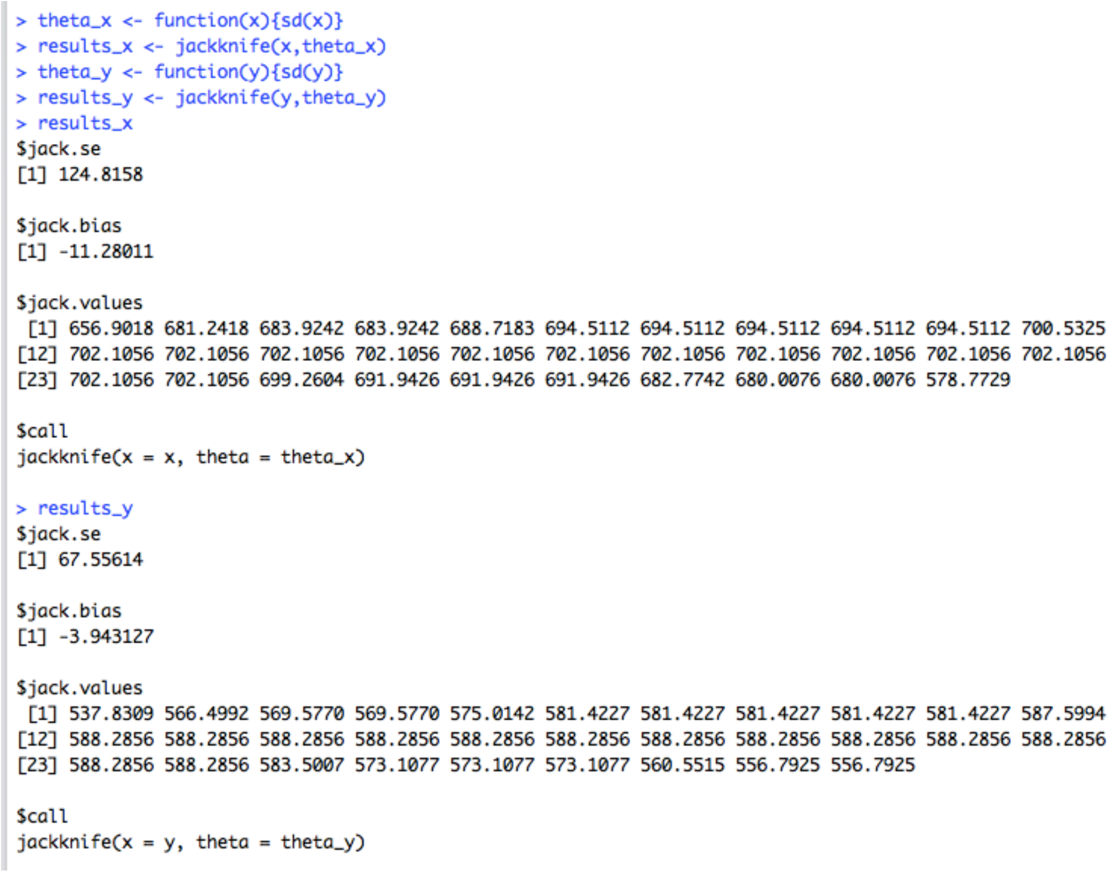


```
> theta_x <- function(x){median(x)}
> results_x <- jackknife(x,theta_x)
> theta_y <- function(y){median(y)}
> results_y <- jackknife(y,theta_y)
> results_x
$jack.se
[1] 0

$jack.bias
[1] 0

$jack.values
 [1] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
[21] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000

$call
jackknife(x = x, theta = theta_x)

> results_y
$jack.se
[1] 0

$jack.bias
[1] 0

$jack.values
 [1] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000
[21] 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000 6000

$call
jackknife(x = y, theta = theta_y)
```

Figure 3.b

Figure 3.b shows that the bias of sample median estimator under the jackknife method is 0, both with outlier and without outlier. The variance of estimator is o with and without outlier.

- Sample Standard Deviation

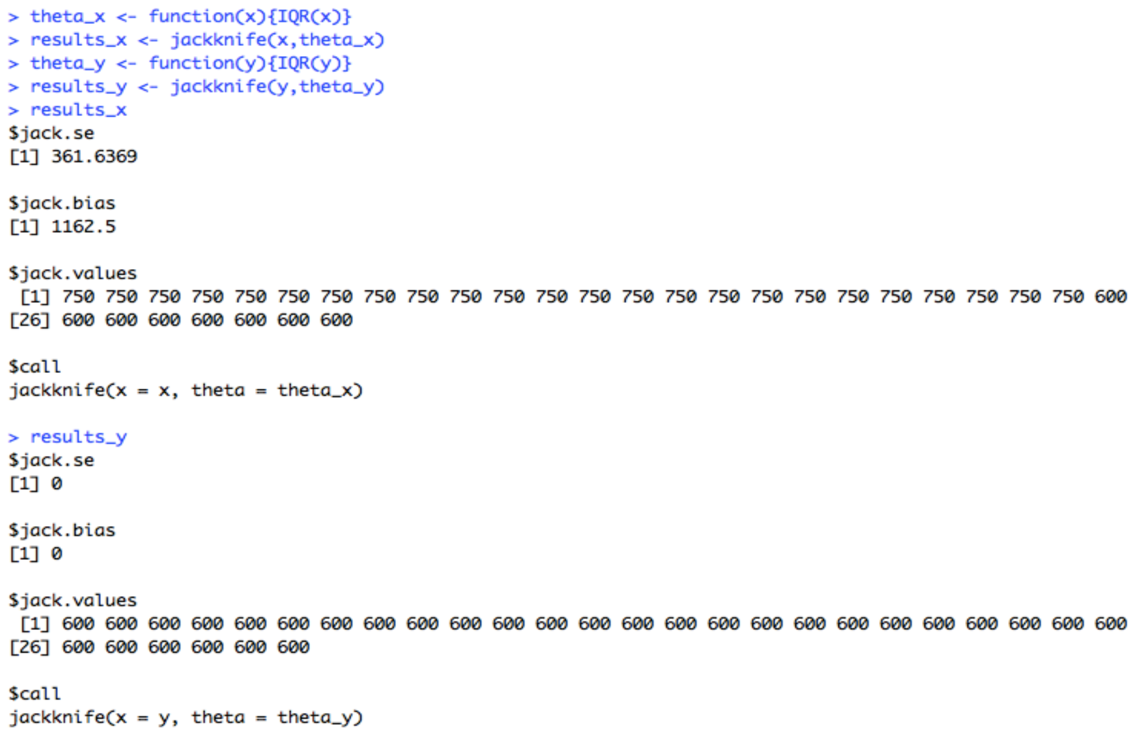


```
> theta_x <- function(x){sd(x)}
> results_x <- jackknife(x,theta_x)
> theta_y <- function(y){sd(y)}
> results_y <- jackknife(y,theta_y)
> results_x
$jack.se
[1] 124.8158

$jack.bias
[1] -11.28011

$jack.values
 [1] 656.9018 681.2418 683.9242 683.9242 688.7183 694.5112 694.5112 694.5112 694.5112 694.5112 700.5325
[12] 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056 702.1056
[23] 702.1056 702.1056 699.2604 691.9426 691.9426 691.9426 682.7742 680.0076 680.0076 578.7729

$call
jackknife(x = x, theta = theta_x)

> results_y
$jack.se
[1] 67.55614

$jack.bias
[1] -3.943127

$jack.values
 [1] 537.8309 566.4992 569.5770 569.5770 575.0142 581.4227 581.4227 581.4227 581.4227 581.4227 587.5994
[12] 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856 588.2856
[23] 588.2856 588.2856 583.5007 573.1077 573.1077 573.1077 560.5515 556.7925 556.7925

$call
jackknife(x = y, theta = theta_y)
```

Figure 3.c

Figure 3.c shows that the bias under the jackknife method is -11.28011with outlier and -3.943127 without outlier. The variance of estimator is (SE)^2=(1224.8158)^2=1500174 with outlier and (67.55164)^2=4563.224 without outlier. The bias and variance are smaller when the outlier is eliminated under the jackknife method.

- IQR



```
> theta_x <- function(x){IQR(x)}
> results_x <- jackknife(x,theta_x)
> theta_y <- function(y){IQR(y)}
> results_y <- jackknife(y,theta_y)
> results_x
$jack.se
[1] 361.6369

$jack.bias
[1] 1162.5

$jack.values
 [1] 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 750 600
[26] 600 600 600 600 600 600 600

$call
jackknife(x = x, theta = theta_x)

> results_y
$jack.se
[1] 0

$jack.bias
[1] 0

$jack.values
 [1] 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600
[26] 600 600 600 600 600 600

$call
jackknife(x = y, theta = theta_y)
```

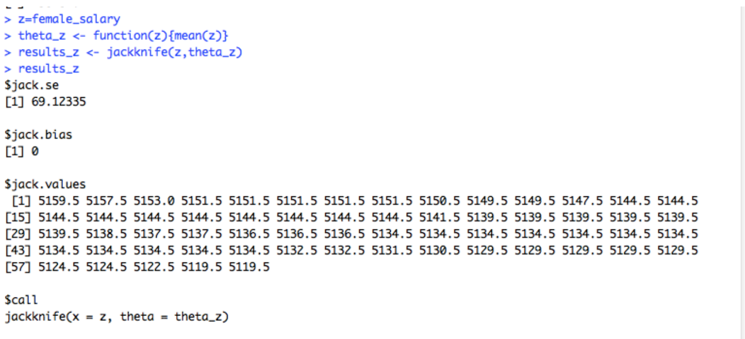
Figure 3.d

Figure 3.d shows that the bias under the jackknife method is 1162.5 with outlier and 0 without outlier. The variance of estimator is (SE)^2=(361.6369)^2=130781.2 with outlier and 0 without outlier. The bias and variance are smaller when the outlier is eliminated under the jackknife method.

- For Female Salary: No outlier in the sample; the same procedure is performed and the result is summarized below. R code on next page.
  - Sample Mean
    - Bias = 0
    - Variance = (69.12335)^2= 4778.038
  - Sample Median
    - Bias = -914.7541
    - Variance= 116.1739^2=13496.38
  - Sample SD
    - Bias= -1.946738
    - Variance = 45.84659^2=2101.91
  - IQR
    - Bias = 0
    - Variance =0

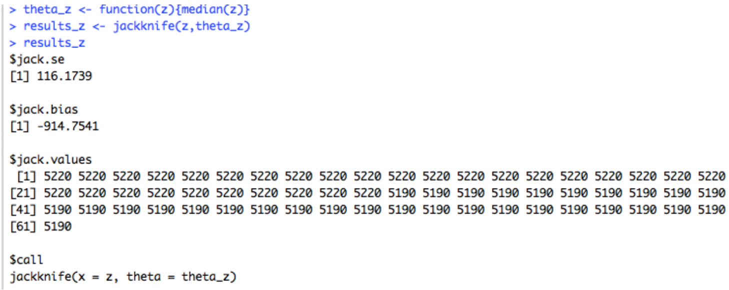
R code

Mean



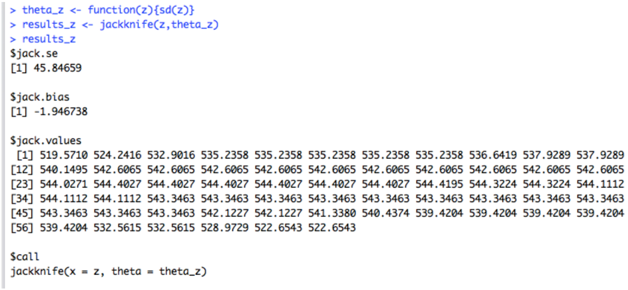
```
> z=female_salary
> theta_z <- function(z){mean(z)}
> results_z <- jackknife(z,theta_z)
> results_z
$jack.se
[1] 69.12335

$jack.bias
[1] 0

$jack.values
 [1] 5159.5 5157.5 5153.0 5151.5 5151.5 5151.5 5151.5 5151.5 5150.5 5149.5 5149.5 5147.5 5144.5 5144.5
[15] 5144.5 5144.5 5144.5 5144.5 5144.5 5144.5 5144.5 5144.5 5141.5 5139.5 5139.5 5139.5 5139.5 5139.5
[29] 5139.5 5138.5 5137.5 5137.5 5136.5 5136.5 5136.5 5134.5 5134.5 5134.5 5134.5 5134.5 5134.5 5134.5
[43] 5134.5 5134.5 5134.5 5134.5 5134.5 5132.5 5132.5 5131.5 5130.5 5129.5 5129.5 5129.5 5129.5 5129.5
[57] 5124.5 5124.5 5122.5 5119.5 5119.5

$call
jackknife(x = z, theta = theta_z)
```

Median



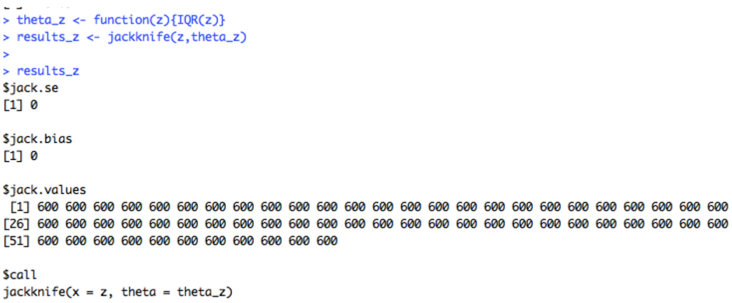
```
> theta_z <- function(z){median(z)}
> results_z <- jackknife(z,theta_z)
> results_z
$jack.se
[1] 116.1739

$jack.bias
[1] -914.7541

$jack.values
 [1] 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220
[21] 5220 5220 5220 5220 5220 5220 5220 5220 5220 5220 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190
[41] 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190 5190
[61] 5190

$call
jackknife(x = z, theta = theta_z)
```

SD



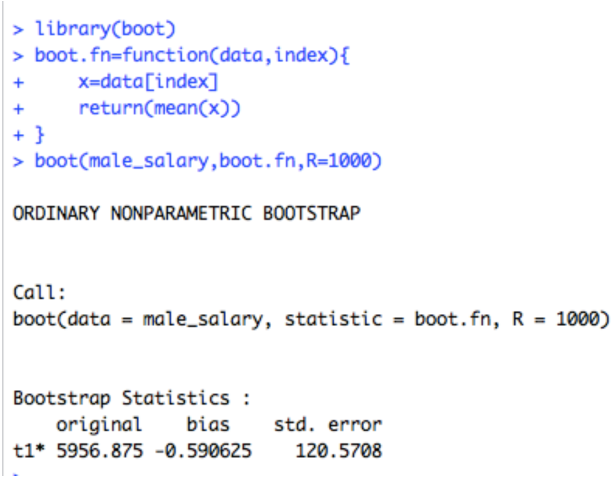
```
> theta_z <- function(z){sd(z)}
> results_z <- jackknife(z,theta_z)
> results_z
$jack.se
[1] 45.84659

$jack.bias
[1] -1.946738

$jack.values
 [1] 519.5710 524.2416 532.9016 535.2358 535.2358 535.2358 535.2358 535.2358 536.6419 537.9289 537.9289
[12] 540.1495 542.6065 542.6065 542.6065 542.6065 542.6065 542.6065 542.6065 542.6065 542.6065 542.6065
[23] 544.0271 544.4027 544.4027 544.4027 544.4027 544.4027 544.4027 544.4195 544.3224 544.3224 544.1112
[34] 544.1112 544.1112 543.3463 543.3463 543.3463 543.3463 543.3463 543.3463 543.3463 543.3463 543.3463
[45] 543.3463 543.3463 543.3463 542.1227 542.1227 541.3380 540.4374 539.4204 539.4204 539.4204 539.4204
[56] 539.4204 532.5615 532.5615 528.9729 522.6543 522.6543

$call
jackknife(x = z, theta = theta_z)
```

IQR



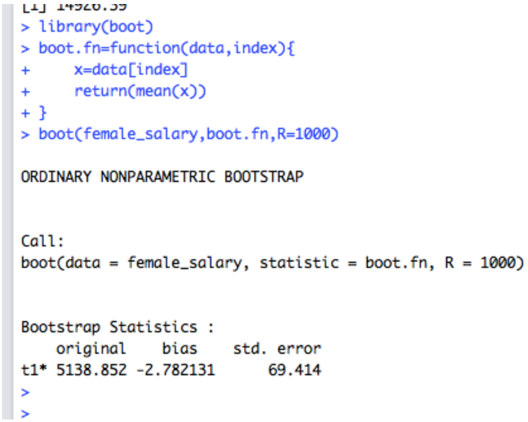
```
> theta_z <- function(z){IQR(z)}
> results_z <- jackknife(z,theta_z)
>
> results_z
$jack.se
[1] 0

$jack.bias
[1] 0

$jack.values
 [1] 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600
[26] 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600 600
[51] 600 600 600 600 600 600 600 600 600 600 600

$call
jackknife(x = z, theta = theta_z)
```

- Bootstrap
    - Sample Mean
        - Male
            - Bias = -0.590625
            - Variance = 120.5708^2=14537.32



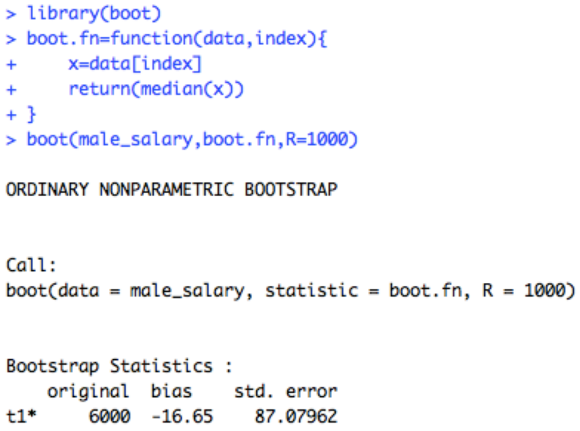
```
> library(boot)
> boot.fn=function(data,index){
+     x=data[index]
+     return(mean(x))
+ }
> boot(male_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = male_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 5956.875 -0.590625    120.5708
```

        - Female
            - Bias=-0.832623
            - Variance =69.47406^2=4826.645



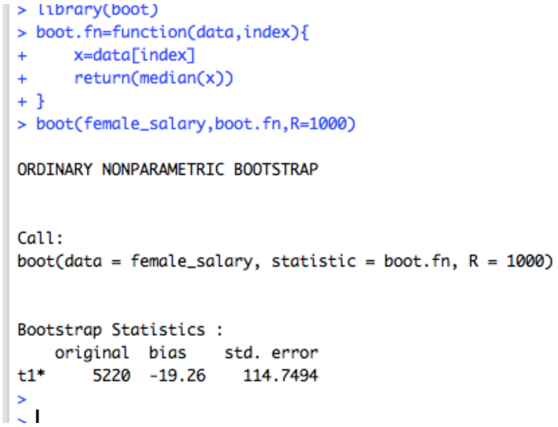
```
[1] 14926.39
> library(boot)
> boot.fn=function(data,index){
+     x=data[index]
+     return(mean(x))
+ }
> boot(female_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = female_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 5138.852 -2.782131      69.414
>
>
```

- Sample Median
    - Male
        - Bias = -16.65
        - Variance =87.07962^2= 7582.86



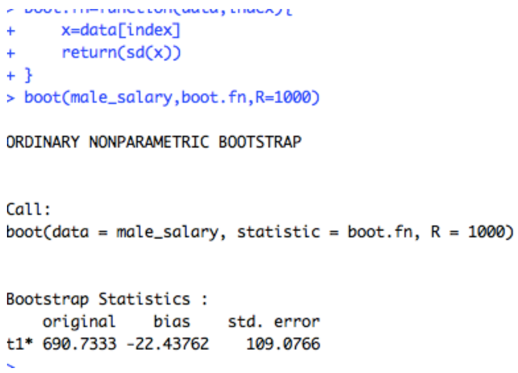
```
> library(boot)
> boot.fn=function(data,index){
+     x=data[index]
+     return(median(x))
+ }
> boot(male_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = male_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*     6000  -16.65    87.07962
```

- Female
    - Bias =-19.26
    - Variance =114.7494^2=13167.42



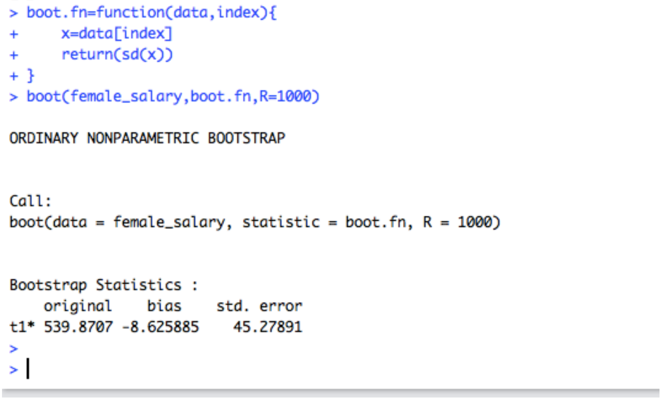
```
> library(boot)
> boot.fn=function(data,index){
+     x=data[index]
+     return(median(x))
+ }
> boot(female_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = female_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*     5220  -19.26    114.7494
>
> |
```

- Sample SD
  - Male
    - Bias =22.43762
    - Variance = 109.0766^2=11897.7



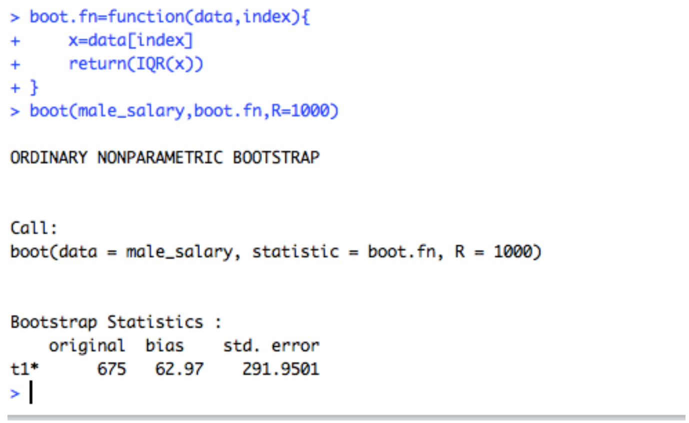
```
> boot.fn=function(data,index){
+     x=data[index]
+     return(sd(x))
+ }
> boot(male_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = male_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 690.7333 -22.43762    109.0766
>
```

  - Female
    - Bias =-8.625885
    - Variance =45.27891^2=2050.18



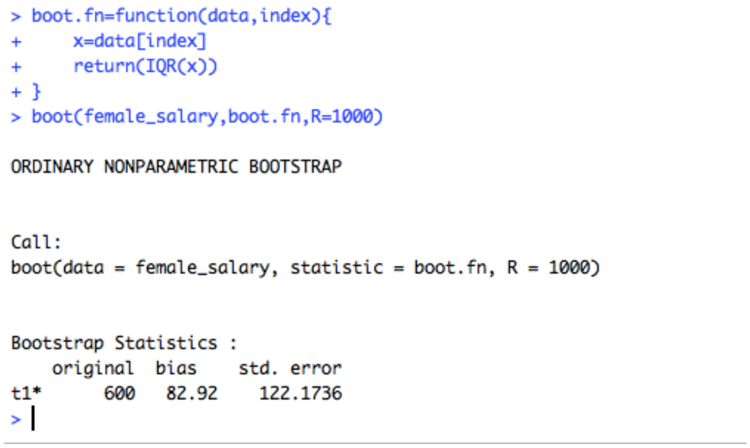
```
> boot.fn=function(data,index){
+     x=data[index]
+     return(sd(x))
+ }
> boot(female_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = female_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 539.8707 -8.625885    45.27891
>
> |
```

- IOR
    - Male
        - Bias= 62.97
        - Variance =291.9501^2=85234.86



```
> boot.fn=function(data,index){
+     x=data[index]
+     return(IQR(x))
+ }
> boot(male_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = male_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*      675   62.97    291.9501
> 
```

-
    - Female
        - Bias=82.92
        - Variance =122.1736^2=14926.39



```
> boot.fn=function(data,index){
+     x=data[index]
+     return(IQR(x))
+ }
> boot(female_salary,boot.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = female_salary, statistic = boot.fn, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*      600   82.92    122.1736
> 
```