Stat 4201 HOMEWORK 1

Mengqi Zong $< mz2326@columbia.edu >$

October 8, 2017

1 Question

Consider the Salary Data (Display 1.3) in Ramsey & Schafer, Chapter 1.

1. Determine whether there are outliers in the combined data, using boxplots.
2. Perform separate EDA, and compute appropriate measures of dispersion for the data in each group (i.e., Males and Females).
3. For each of the estimates computed in 2 above, determine the bias and variance using each of the following methods:
   - Jackknife
   - Bootstrap

2 Answers

2.1 Question 1

The boxplot is show in Fig **??**. It shows that there is one outlier in the combined data, it is "8100 MALE".

2.2 Question 2

2.2.1 EDA

I use histograms, stem-and-leaf diagrams, and box plots. The histogram for males is shown in Fig **??**, and the histogram for females is shown in Fig-**??**. The box plot for males is shown in Fig-**??**, and the box plot for females is shown in Fig-**??**. As to the stem-and-leaf diagrams, the two diagrams are listed below.

The stem-and-leaf diagram for males is:

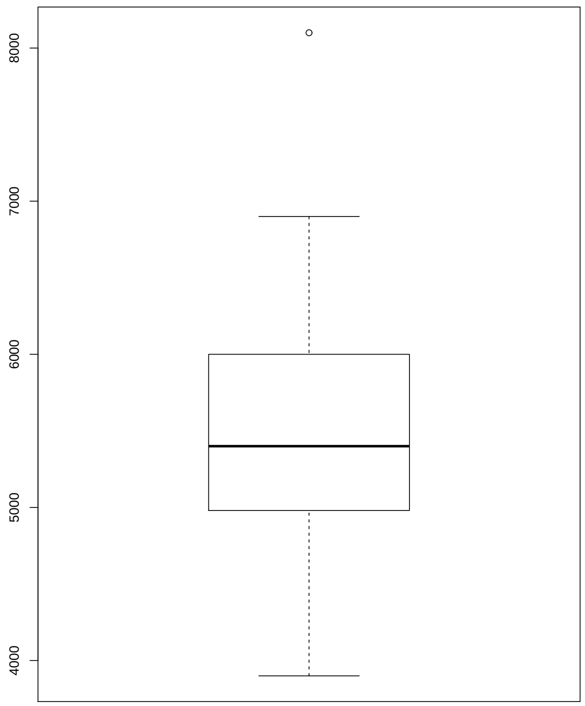

Figure 1: Boxplot for the combined data

```
The decimal point is 3 digit(s) to the right of the |

4 | 6
5 | 011244444
5 | 7
6 | 00000000000003
6 | 666899
7 |
7 |
8 | 1
```

The stem-and-leaf diagram for females is:

```
The decimal point is 3 digit(s) to the right of the |
4 | 6
5 | 011244444
5 | 7
6 | 00000000000003
6 | 666899
7 |
7 |
8 | 1
```

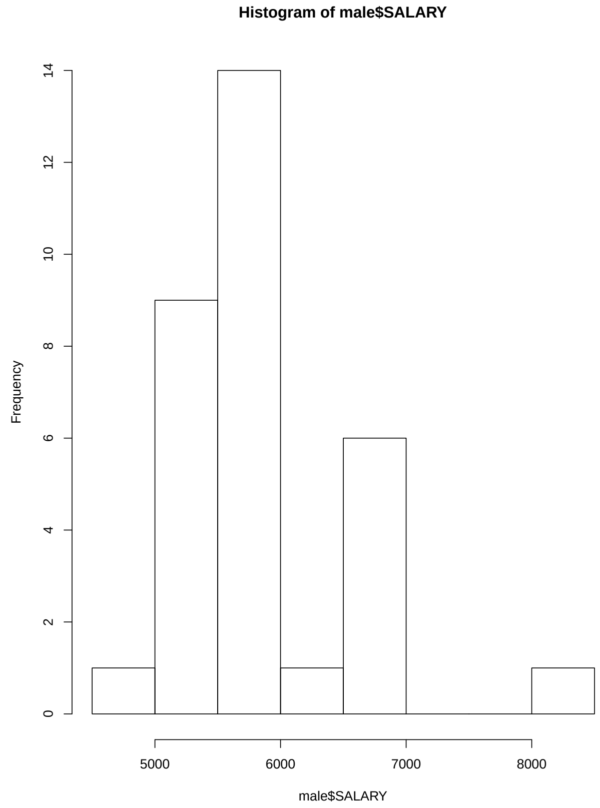

Figure 2: Histogram for male salaries

2.2.2 Measures of Dispersions

I use standard deviation and interquartile range to measure the dispersions. The calculated data is show in Table **??**.

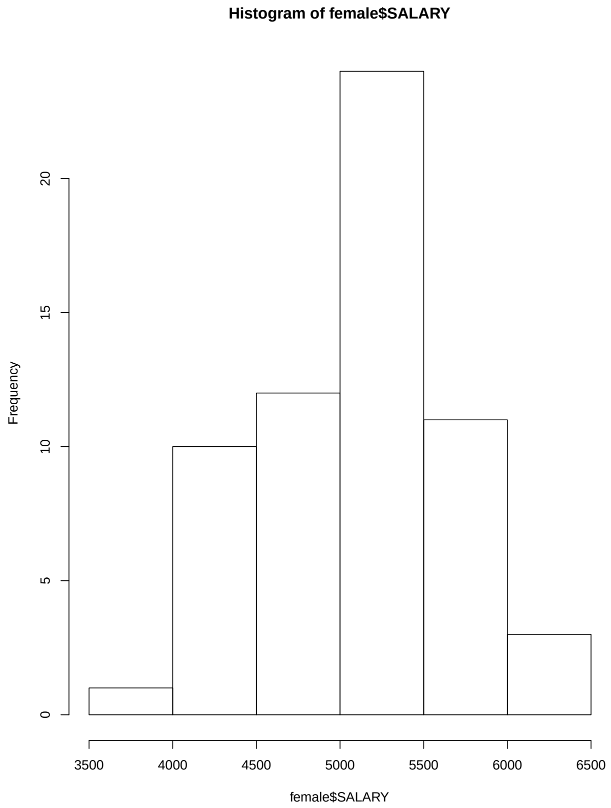

Figure 3: Histogram for female salaries

	Standard Deviation	Interquartile Range
Males	690.7333	675
Females	539.8707	600

Table 1: Measures of Dispersions

2.3 Question 3

The summary of the estimated bias, variance is shown in Table **??**.

3 Appendix

The code is listed below:

```
# Advanced Data Analysis Homework 1
```

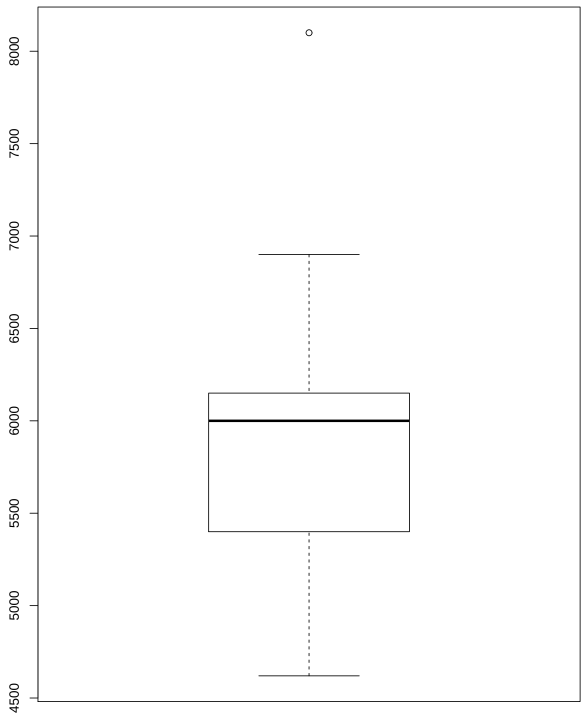

Figure 4: Boxplot for male salaries

Method	Variable	SD (M)	IQL (M)	SD (F)	IQL (F)
Jackknife	Bias	-11.28011	1162.5	-1.946738	0
	Standard Error	124.8158	361.6369	45.84659	0
Bootstrap	Bias	-13.58164	46.47	-9.533073	77.76
	Standard Error	117.2164	302.896	46.26237	126.0459

Table 2: Measures of Dispersions

```
#
# Name Mengqi Zong
# UNI: mz2326
#

library(boot)
library(bootstrap)
```

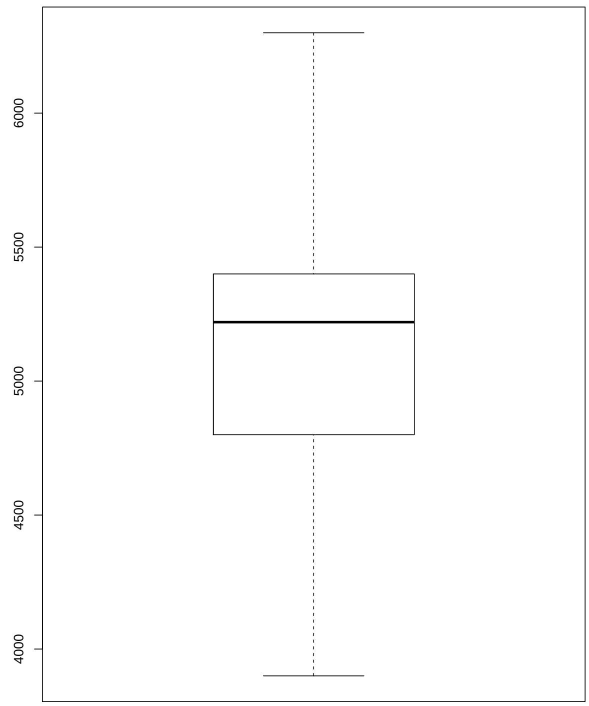

Figure 5: Boxplot for male salaries

```
male <- read.table("male.data", header=TRUE)
female <- read.table("female.data", header=TRUE)
# "mix.data" is the combination of "male.data" and "female.data"
mix <- read.table("mix.data", header=TRUE)

#######################
##  problem 1
#######################

postscript(file="~/Documents/LaTeX/stat4201-hmwk1/boxplot_mix.eps", onefile=FALSE, horizonta
boxplot(mix$SALARY)
dev.off()

#######################
```

```
##  problem 2
#######################

# male - EPA
postscript(file="~/Documents/LaTeX/stat4201-hmwk1/boxplot_male.eps", onefile=FALSE, horizont
boxplot(male$SALARY)
dev.off()

stem(male$SALARY)

postscript(file="~/Documents/LaTeX/stat4201-hmwk1/hist_male.eps", onefile=FALSE, horizontal=
hist(male$SALARY)
dev.off()

# male - Standard Deviation
sd.male <- sd(male$SALARY)
cat("Male Salary Standard Deviation = ")
print(sd.male)

# male - IQR
iqr.male <- IQR(male$SALARY)
cat("Male IQR = ")
print(iqr.male)

# female - EPA
postscript(file="~/Documents/LaTeX/stat4201-hmwk1/boxplot_female.eps", onefile=FALSE, horizo
boxplot(female$SALARY)
dev.off()

stem(female$SALARY)

postscript(file="~/Documents/LaTeX/stat4201-hmwk1/hist_female.eps", onefile=FALSE, horizonta
hist(female$SALARY)
dev.off()

# female - Standard Deviation
sd.female <- sd(female$SALARY)
cat("Female Salary Standard Deviation = ")
print(sd.female)

# female - IQR
iqr.female <- IQR(female$SALARY)
cat("Female IQR = ")
print(iqr.female)

#######################
```

```
##  problem 3
#######################

# male - Jackknife Standard Deviation
jacksd.male <- jackknife(male$SALARY, sd)
cat("Male Salary Standard Deviation Using Jackknife = ")
print(jacksd.male)

# male - Bootstrap Standard Deviation
foosd <- function(d, i) {
  d2 <- d[i,]
  return(sd(d2$SALARY))
}
bootsd.male <- boot(male, foosd, R = 500)
cat("Male Salary Standard Deviation Using Bootstrap = ")
print(bootsd.male)

# male - Jackknife IQR
jackiqr.male <- jackknife(male$SALARY, IQR)
cat("Male Salary IQR Using Jackknife = ")
print(jackiqr.male)

# male - Bootstrap IQR
fooiqr <- function(d, i) {
  d2 <- d[i,]
  return(IQR(d2$SALARY))
}
bootiqr.male <- boot(male, fooiqr, R = 500)
cat("Female Salary IQR Using Bootstrap = ")
print(bootiqr.male)

# female - Jackknife Standard Deviation
jacksd.female <- jackknife(female$SALARY, sd)
cat("Female Salary Standard Deviation Using Jackknife = ")
print(jacksd.female)

# female - Bootstrap Standard Deviation
bootsd.female <- boot(female, foosd, R = 500)
cat("Female Salary Standard Deviation Using Bootstrap = ")
print(bootsd.female)

# female - Jackknife IQR
jackiqr.female <- jackknife(female$SALARY, IQR)
cat("FEMale Salary IQR Using Jackknife = ")
print(jackiqr.female)
```

```
# female - Bootstrap IQR
fooiqr <- function(d, i) {
  d2 <- d[i,]
  return(IQR(d2$SALARY))
}
bootiqr.female <- boot(female, fooiqr, R = 500)
cat("Female Salary IQR Using Bootstrap = ")
print(bootiqr.female)
```