STA 303H1S / 1002 HS -Winter 2018 Assignment # 1 Solutions

“In or Out”

Shivon Sue-Chee

January 26, 2018

1. (30 marks) **Modified box plots versus Skeletal box plots in R.**

   (a) (10 marks) For Spock’s Judge, the first and third quartiles were 13.30% and 17.70% respectively. Hence, using the 1.5IQR Rule, the upper and lower fences were:

   Upper fence (SPOCKS): 17.70 +1.5 (17.70-13.30)=24.3% and
   Lower fence (SPOCKS): 13.30- 1.5 (17.70-13.30)= 6.7%.

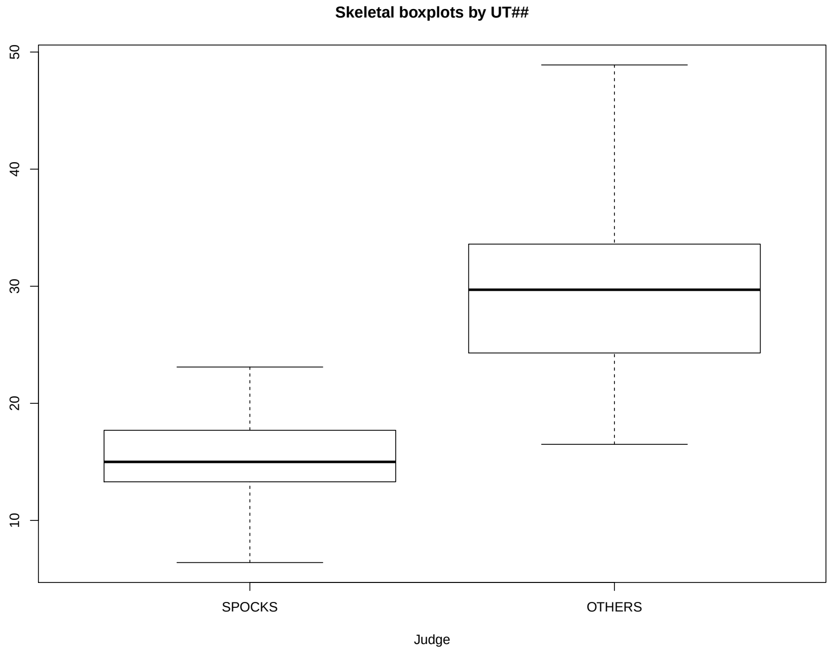
   The only point beyond these fences was 6.4%. Therefore, this point was identified as an outlier in the lower tail of the data.

   For the other judges, the first and third quartiles were 24.30% and 33.60% respectively. Hence, the upper and lower fences were:

   Upper fence (OTHERS): 33.60 +1.5 (33.60-24.30)=47.55% and
   Lower fence (OTHERS): 24.30- 1.5 (33.60-24.30)= 10.35%.

   The only point beyond these fences was 48.9%. Therefore, this point was identified as an outlier in the upper tail of the data.

   (b) (15 marks) Setting the `option, range=0` in the *R* function `boxplot()` to hide outliers, the side-by-side box plots of percent of women on venires for Spock’s judge and the other judges were recreated.



(c) (5 marks) It is useful to identify potential outliers if they exist in the data. For that reason, modified box plots are more useful than skeletal box plots. However, if data are roughly normal, the two types of box plot visualizations would be the same.

2. (70 marks) **Heights of males and females in our class with and without an (potential influential point) observation.**

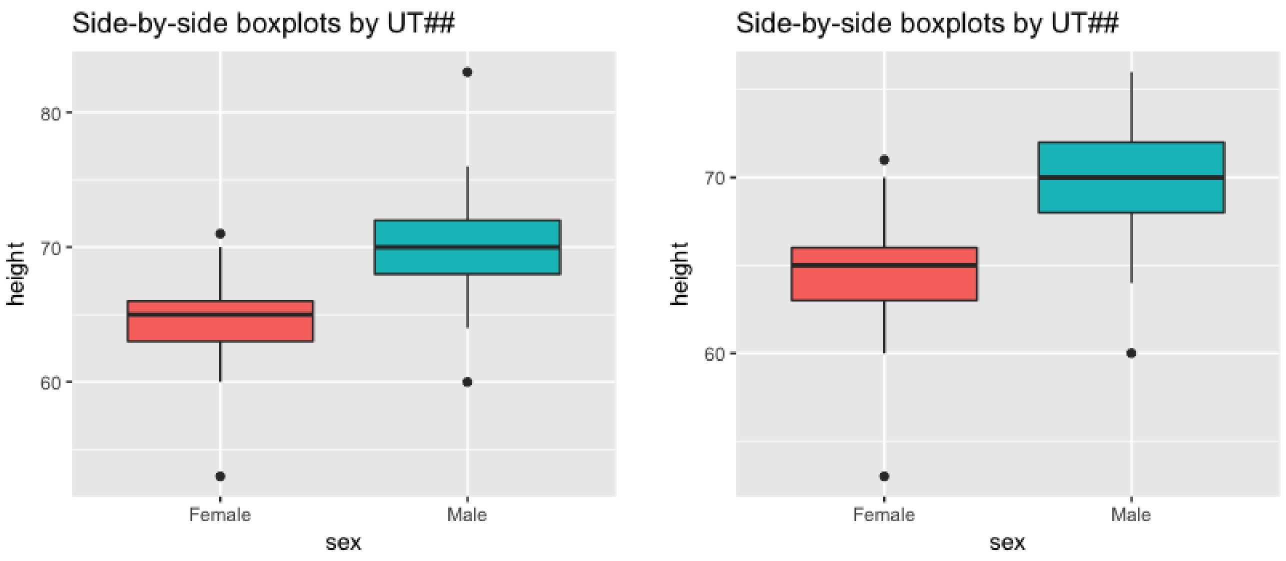
(a) (5 marks) The data were based on an online survey which is a type of observational study. A treatment was not assigned in the study of heights of students in our class. In fact, it was impossible to assign students to a certain 'sex'. Since this was not an experiment, any significant association between 'sex' and 'height' cannot be deemed as a cause-effect relationship.

(b) (5 marks) The categorical variables were 'sex' and 'id'. Both describe characteristic qualities of students; 'id' identifies individual students and 'sex' is an indicator of their biological nature. Both variables are not numerical measurements; in other words, there is no numeric difference from one level to the other level of the same variable.

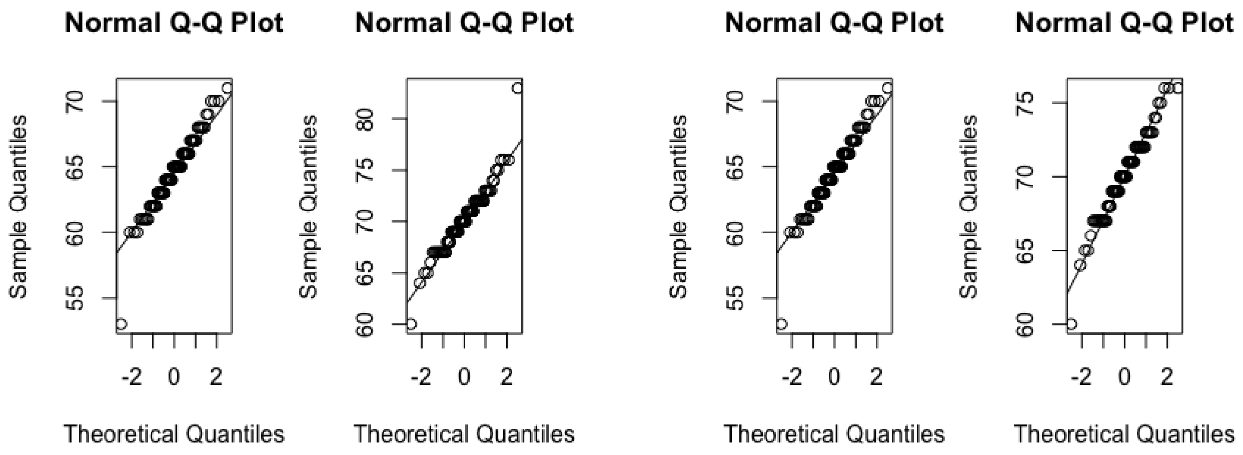
Categorical variable	Number of Levels	Names of Levels
`sex`	2	Males, Females
`id`	166	1, 2, 3, 4, 5, . . ., 166

(c) (e) (20+25 marks) Comparison heights of *Males* and *Females* with and without an observation is done via the pooled two-sample t-test and results are given in the table below. (i) Side-Beside box plots and Normal QQ plots are given on the next page.

Parts	(c) With observation	(e) Without observation
ii. Hypotheses	$H_0 : \mu_{females} - \mu_{males} = 0,\ H_a : \mu_{females} - \mu_{males} \neq 0$	
iii. Test statistic	$-12.08 \sim t_{164}$	$-12.37 \sim t_{163}$
iv. Assumptions	1. Two independent samples from Normal populations 2. Populations have equal variances.	
v. Diagnostics	1. QQ plots below show that samples are Normal except for a few potential outliers. 2. Equal variances assumption satisfied by F-test. (see p=0.2471)	(see p=0.8644)
vi. P-value	$p \approx 0$	$p \approx 0$
vii. Results	Evidence that males are taller than females.	



i.



v.

(d) (5 marks) Two additional equivalent methods are:

1. Simple linear regression with 1 dummy variable
2. One-way Analysis of Variance with 2 groups

(e) See relevant codes in Appendix and analyses in part (c) above.

(f) (10 marks) Though the 32*nd* observation removed was a identified as an extremely large height among the males, our results did not change dramatically from parts (c) to (e). Hence the point was not influential.

Appendix of R code and Output

```
> #Question 1- Juries data
> library(Sleuth3)
> jury<-case0502
> attach(jury)
The following objects are masked from jury (pos = 4):

    Judge, Percent

> SPOCKS<-Percent[Judge=="Spock's"]
> OTHERS<-Percent[Judge!="Spock's"]
> boxplot(SPOCKS, OTHERS,range=0,xlab="Judge",names=c("SPOCKS","OTHERS"),
main="Skeletal boxplots by UT##")
> summary(SPOCKS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.40   13.30   15.00   14.62   17.70   23.10
> summary(OTHERS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.50   24.30   29.70   29.49   33.60   48.90

> #using 1.5 IQR RULE for SPOCKS
> sum_sp=summary(SPOCKS)
> uf_spock=sum_sp[5]+1.5*(sum_sp[5]-sum_sp[2])
> lf_spock=sum_sp[2]-1.5*(sum_sp[5]-sum_sp[2])
> uf_spock
3rd Qu.
   24.3
> lf_spock
1st Qu.
    6.7
> SPOCKS[SPOCKS>uf_spock]
numeric(0)
> SPOCKS[SPOCKS<lf_spock]
[1] 6.4
>
> #using 1.5 IQR RULE for OTHERS
> sum_ot=summary(OTHERS)
> uf_other=sum_ot[5]+1.5*(sum_ot[5]-sum_ot[2])
> lf_other=sum_ot[2]-1.5*(sum_ot[5]-sum_ot[2])
> uf_other
3rd Qu.
  47.55
> lf_other
1st Qu.
  10.35
> OTHERS[OTHERS>uf_other]
```

```
[1] 48.9
> OTHERS[OTHERS<lf_other]
numeric(0)
>
> #Question 2- Class Heights
> shivon.hdata<-read.csv("assign1data.csv", header=T)
> attach(shivon.hdata)
> library(ggplot2)
> box2c=ggplot(shivon.hdata, aes(x=sex,y=height, fill=sex))+geom_boxplot()
> print(box2c+ggtitle("Side-by-side boxplots by UT##")+theme(legend.position="none"))
> #boxplot(height˜sex, ylab="Height(cm)",names=c("Females","Males"),
main="Side-by-side boxplots by UT##")
>
> var.test(height˜sex)

F test to compare two variances

data:  height by sex
F = 0.77418, num df = 85, denom df = 79, p-value = 0.2471
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4996117 1.1954099
sample estimates:
ratio of variances
         0.7741754

> t.test(height˜sex,var.equal=T )

Two Sample t-test

data:  height by sex
t = -12.076, df = 164, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.525811 -4.691630
sample estimates:
mean in group Female   mean in group Male
            64.61628             70.22500

>
> par(mfrow=c(1,2))
> qqnorm(height[sex=="Female"])
> qqline(height[sex=="Female"])
> qqnorm(height[sex=="Male"])
> qqline(height[sex=="Male"])
>
> shivon.subset<-shivon.hdata[-32,]
> par(mfrow=c(1,1))
> box2e=ggplot(shivon.subset, aes(x=s32,y=h32, fill=s32))+geom_boxplot()
> print(box2e+labs(title="Side-by-side boxplots by UT##",x="sex",y="height")
+theme(legend.position="none"))
> #boxplot(height~sex,data=shivon.subset, ylab="Height(cm)",names=c("Females","Males"),
main="Side-by-side boxplots by UT##")
> var.test(height~sex, data=shivon.subset)

F test to compare two variances

data:  height by sex
F = 0.96341, num df = 85, denom df = 78, p-value = 0.8644
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6206248 1.4893223
sample estimates:
ratio of variances
          0.963413

> t.test(height~sex,var.equal=T, data=shivon.subset)

Two Sample t-test

data:  height by sex
t = -12.372, df = 163, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.316397 -4.577627
sample estimates:
mean in group Female   mean in group Male
            64.61628             70.06329


> par(mfrow=c(1,2))
> h32<-shivon.subset$height
> s32<-shivon.subset$sex
> qqnorm(h32[s32=="Female"])
> qqline(h32[s32=="Female"])
> qqnorm(h32[s32=="Male"])
> qqline(h32[s32=="Male"])
```

Grading Scheme Notes

1. (30 marks)

   (a) (10marks): -5 if little or no work done

   (b) (15 marks): -5marks: No unique labeling (using last 4 digits of student id) of boxplot; -5marks: If outliers were removed(using outline=F) instead of range=0 (refer to code)

   (c) (5 marks) -5 if no work done

2. (70 marks)

   (a) (5 marks) -5 if incorrect

   (b) (5 marks) -5 if no work done.

   (c) (20 marks) -5marks for irrelevant R (text and numbers) output here and throughout the rest of the main part. Only relevant numbers should be quoted. Full output should be in the appendix!
   -5marks for no checks for normality (via residual plots and/or qqplots) for 2c or 2e.

   (d) (5 marks) -5 if both answers are wrong; Method used in part c already should not be repeated. Any two of (two-sample t test, simple linear regression with dummy variable, one -way aNOVA) make up the answer.

   (e) (25 marks) -5: See R code for correct id removed

   (f) (10 marks) -10 if no work done