stat4051hw3

Mingming Xu

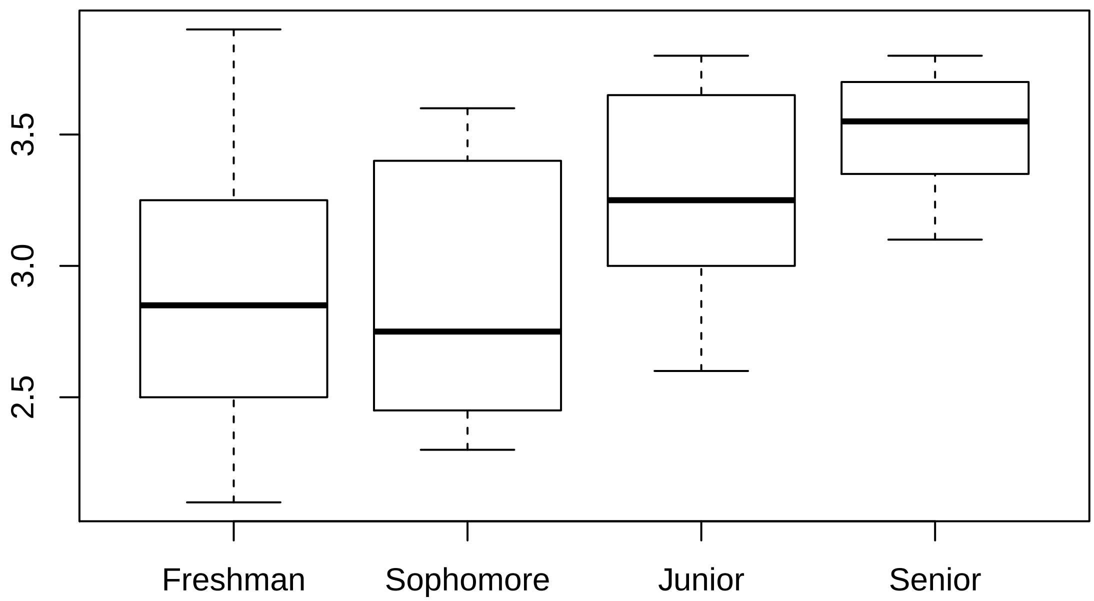
2019/10/5

1.

```
c_f=c(2.8,2.1,2.7,3.0,2.3,2.9,3.5,3.9)
c_so=c(2.5,2.3,2.9,3.5,2.6,2.4,3.3,3.6)
c_j=c(3.1,2.9,3.2,3.8,2.6,3.6,3.3,3.7)
c_se=c(3.8,3.6,3.5,3.1,3.2,3.5,3.8,3.6)
GPAdata=data.frame(Freshman=c_f,Sophomore=c_so,Junior=c_j,Senior=c_se)
GPA=stack(GPAdata)
```

a.

```
boxplot(GPAdata)
```



Difference between sample means is small in comparison to variability within group.

b.

```
model.aov = aov(GPA$values~GPA$ind)
summary(model.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## GPA$ind      3  2.226  0.7421   3.502 0.0283 *
## Residuals   28  5.932  0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At $\alpha = 0.5$,because the p-value is larger than α,we could not say that there is statistical difference among the class years.At $\alpha = 0.01$,because the p-value is less than α,we could say there is statistical difference among the class years.

Yes,my results support my graphic explanation from part a.

c.

```
mu_g=tapply(GPA$values,GPA$ind,mean)
mu=mean(GPA$values)
mu_g;mu
```

```
##  Freshman Sophomore    Junior    Senior
##    2.9000    2.8875    3.2750    3.5125
```

```
## [1] 3.14375
```

```
SST=sum((GPA$values-mu)^2)
SSG=sum((mu_g-mu)^2)
SSE=sum((GPAdata$Freshman-mu_g[1])^2)+sum((GPAdata$Sophomore-mu_g[2])^2)+sum((GPAdata$Junior-mu_g[3])^2
SSG;SSE;SSG+SSE;SST
```

```
## [1] 0.2782813
```

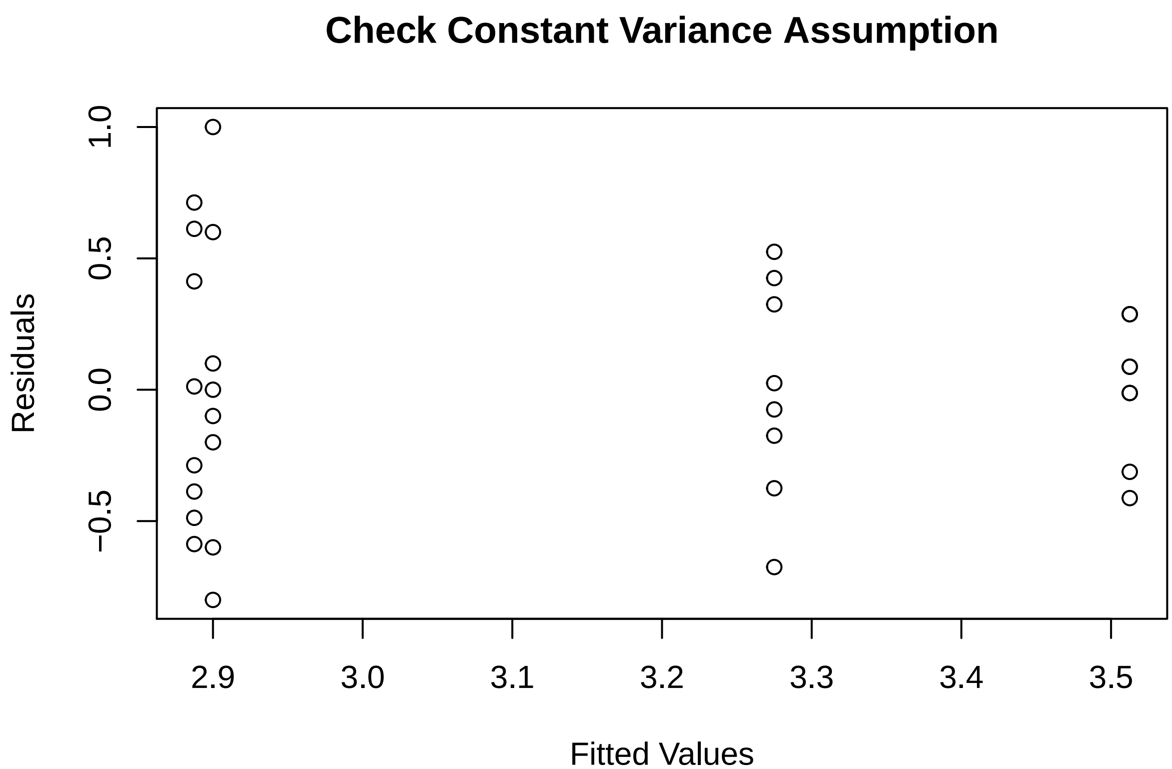
```
## [1] 5.9325
```

```
## [1] 6.210781
```

```
## [1] 8.15875
```

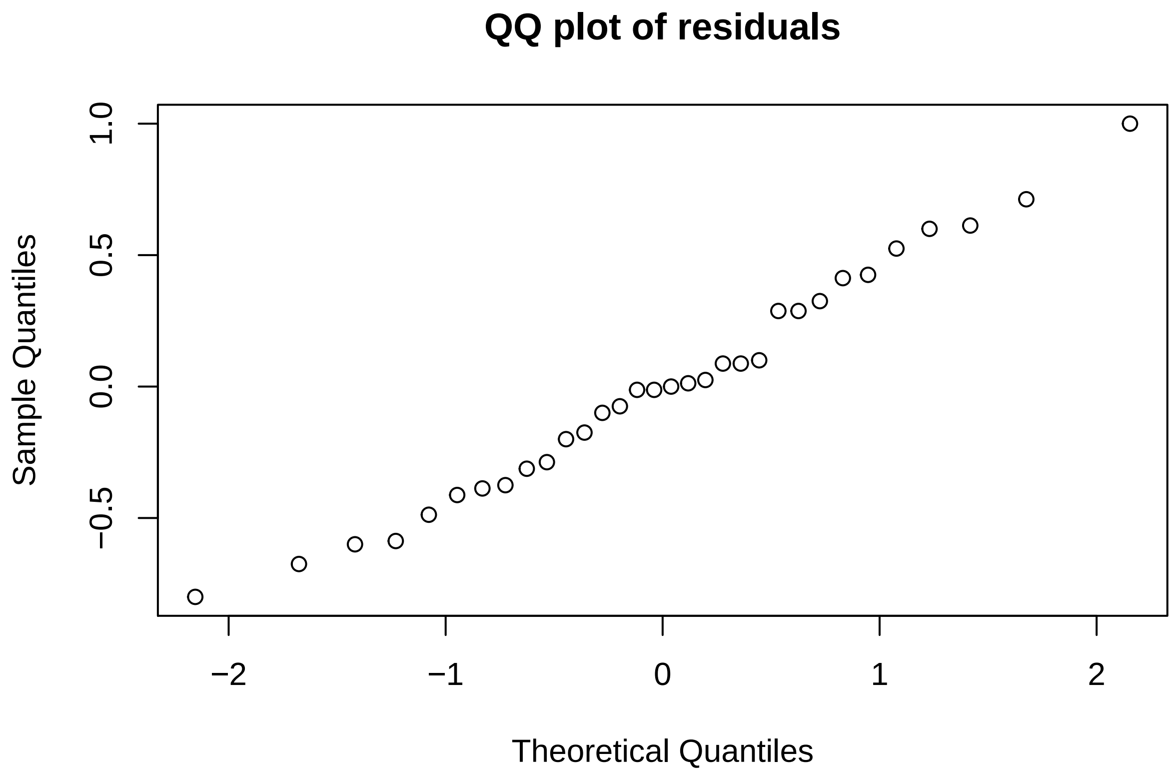
d.

```
##test assumptions:
##Independece:Because these students from each class were randomly selected,these observations are inde
##Constant variance:
residuals=model.aov$residuals
fitted.values=model.aov$fitted.values
plot(fitted.values,residuals,ylab="Residuals",xlab="Fitted Values",main= "Check Constant Variance Assum
```




```
##abline function can not be run in R
##abline(h=0,col=c("red"))
```

```
##Because there isno any discernible pattern in the plot, variance is constant.
##Normality:
qqnorm(residuals,main = "QQ plot of residuals")
```




```
##qqline function can not be run in R
##qqline(residuals)
## From the plot, we could see that the most of points are on the line.So, it is normality.
```

e.

```
model.lm=lm(GPA$values~GPA$ind)
summary(model.lm)
```

```
##
## Call:
## lm(formula = GPA$values ~ GPA$ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80000 -0.32812 -0.00625  0.29687  1.00000
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.9000     0.1627  17.820   <2e-16 ***
## GPA$indSophomore   -0.0125     0.2301  -0.054   0.9571
## GPA$indJunior       0.3750     0.2301   1.629   0.1144
## GPA$indSenior       0.6125     0.2301   2.661   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4603 on 28 degrees of freedom
## Multiple R-squared:  0.2729, Adjusted R-squared:  0.195
## F-statistic: 3.502 on 3 and 28 DF,  p-value: 0.02831
```

Each t-test:

GPA$indSophomore:It tests if there is diffrence between the mean of Freshman GPA and the mean of Sophomore GPA.

GPA$indJunior:It tests if there is diffrence between the mean of Freshman GPA and the mean of Junior GPA.

GPA$indSenior:It tests if there is diffrence between the mean of Freshman GPA and the mean of Senior GPA.

f.

```r
##l=(1)mu_1+(-1)mu_4
w1=c(1,0,0,-1)
l1_hat=sum(w1*mu_g)
mse=0.2119
se=sqrt(mse)*sqrt(sum((w1^2)/8))
t=l1_hat/se
p_val=2*pt(t,28,lower.tail = FALSE)
l1_hat;t;p_val
```

```
## [1] -0.6125

## [1] -2.661158

## [1] 1.987252
```

Because the p-value is larger than $\alpha = 0.05$,we fail to reject the null hypotheses that GPA of frenshmen is not diffrent from GPA of seniors.

g.

```r
##l=(1/2)mu_1+(-1/2)mu_4
w2=c(1/2,0,0,-1/2)
l2_hat=sum(w2*mu_g)
mse=0.2119
se=sqrt(mse)*sqrt(sum((w2^2)/8))
t=l2_hat/se
p_val=2*pt(t,28,lower.tail = FALSE)
l2_hat;t;p_val
```

```
## [1] -0.30625

## [1] -2.661158

## [1] 1.987252
```

Compared to part(f), the estimate of contrast is changed but t-values and p-value are same to the results in part(g).

h.

```r
##l=(1)mu_3+(-1)mu_4
w3=c(0,0,1,-1)
l3_hat=sum(w3*mu_g)
mse=0.2119
se=sqrt(mse)*sqrt(sum((w3^2)/8))
t=l3_hat/se
```

```r
p_val=2*pt(t,28,lower.tail = FALSE)
l3_hat;t;p_val
```

```
## [1] -0.2375
## [1] -1.031877
## [1] 1.689039
```

Because the p-value is larger than $\alpha = 0.05$,we fail to reject the null hypotheses that GPA of Junior is not diffrent from GPA of seniors.

i.

```r
##l=(-3)mu_1+(-1)mu_2+(1)mu_3+(3)mu_4
w4=c(-3,-1,1,3)
l4_hat=sum(w4*mu_g)
```

j.

```r
##95% Confidence interval:
mse=0.2119
se=sqrt(mse)*sqrt(sum((w4^2)/8))
l4_hat+c(-1,1)*qt(1-0.05/2,28)*se
```

```
## [1] 0.7340888 3.7159112
```

k.

```r
t=l2_hat/se
p_val=2*pt(t,28,lower.tail = FALSE)
t;p_val
```

```
## [1] -0.420766
## [1] 1.322862
```

Because the p-value is larger than $\alpha = 0.05$,we do not enough evidence to reject the null hypothesis that GAP does's increase with class year.

l.

```r
## 3 orthogonal contrasts
c1=c(1,-1/3,-1/3,-1/3)
c2=c(0,1,-1/2,-1/2)
c3=c(0,0,-1,1)

sum(c1*c2);sum(c1*c3);sum(c2*c3)
```

```
## [1] 0
## [1] 0
## [1] 0
```

m.

```r
ss_c1=(sum(c1*mu_g))^2/(sum((c1^2)/8))
ss_c2=(sum(c2*mu_g))^2/(sum((c2^2)/8))
ss_c3=(sum(c3*mu_g))^2/(sum((c3^2)/8))
ss_c1+ss_c2+ss_c3
```

```
## [1] 2.22625
```

```
## which is equal to SSG 2.226.
```

n.

```
TukeyHSD(model.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = GPA$values ~ GPA$ind)
##
## $`GPA$ind`
##                         diff          lwr       upr     p adj
## Sophomore-Freshman -0.0125 -0.640879593 0.6158796 0.9999410
## Junior-Freshman     0.3750 -0.253379593 1.0033796 0.3791406
## Senior-Freshman     0.6125 -0.015879593 1.2408796 0.0581213
## Junior-Sophomore    0.3875 -0.240879593 1.0158796 0.3508591
## Senior-Sophomore    0.6250 -0.003379593 1.2533796 0.0516384
## Senior-Junior       0.2375 -0.390879593 0.8658796 0.7323952
```

o.

```
(mu_g[1]-mu_g[4])+c(-1,1)*qtukey(1-0.05,4,28)/sqrt(2)*sqrt(mse)*sqrt(2/8)
```

```
## [1] -1.24091666  0.01591666
```

This confidence intervals contains 0.

2.

5-step test:

Assumptions:

For this test, these groups and observations are independent.

Hypothesis:

ClassYear: H_0 : All class years are the same H_a : At least one class year is diffrent

Sex: H_0 : The GPA of male is same to the GPA of femlae H_a : The GPA of male is diffrent to the GPA of male.

Class Year and Sex interaction: H_0 : All interactions are the same H_a : At least one interaction is diffrent

Test Statistic:

```
GPASEX=data.frame(classyear=c(rep("Freshman",8),rep("Sophomore",8),rep("Junior",8),rep("Senor",8)),sex=c
summary(aov(gpa~classyear*sex,data = GPASEX))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## classyear      3  2.226  0.7421   3.320 0.0368 *
## sex            1  0.281  0.2813   1.258 0.2731
## classyear:sex  3  0.286  0.0954   0.427 0.7356
## Residuals     24  5.365  0.2235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-values:

Class Year: The p-value is 0.0368.

Sex: The p-value is 0.2731.

Class Year and Sex interaction: The p-value is 0.7356 .

Conclusion:

Class Year: The p-value is 0.0368 which is samller than $\alpha = 0.05$, so we have enough evidence to reject the null hypothesis that all class years are the same.

Sex: The p-value is 0.2731 which is larger than $\alpha = 0.05$, so we do not have enough evidence to reject the null hypothesis that the GPA of male is same to the GPA of femlae.

Class Year and Sex interaction: The p-value is 0.7356 which is larger than $\alpha = 0.05$, so we do not have enough evidence to reject the null hypothesis that all interactions are the same.

Hence, we could say thah the class year can effect GPA

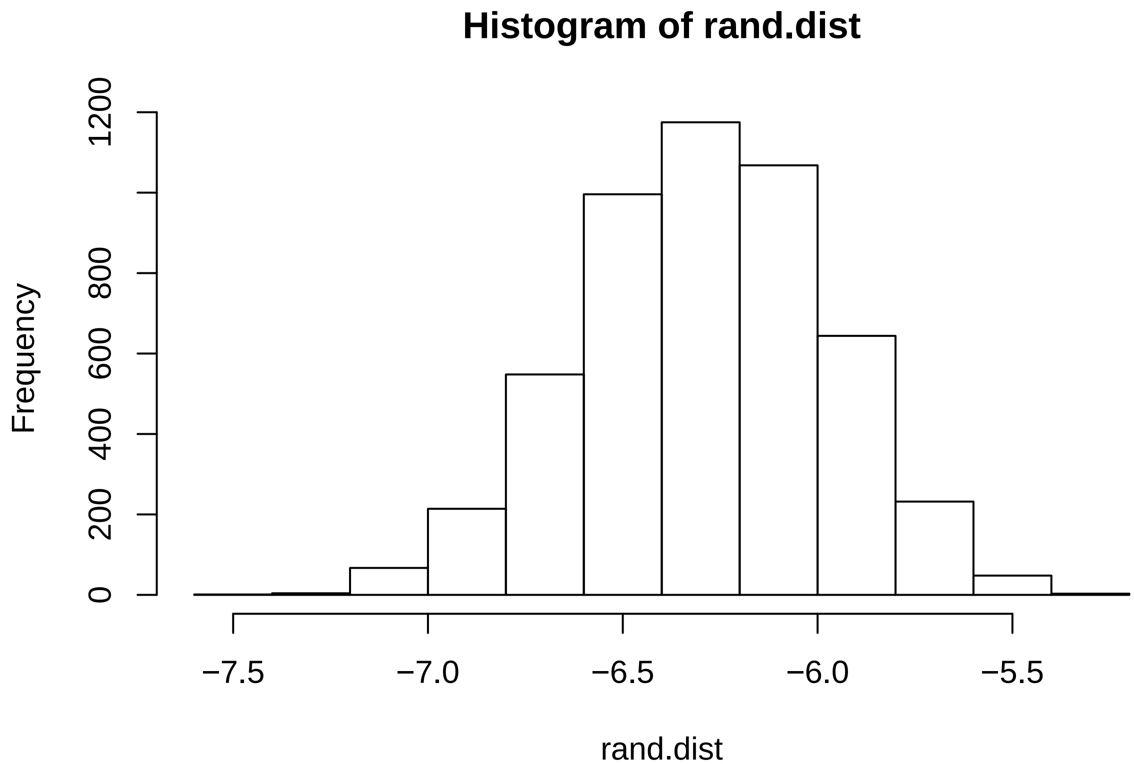
Bonus:

H_0: $\mu_f = \mu_{so} = \mu_j = \mu_{se}$(each class year's mean of GPA is same)

H_a: at least one class year's mean of GPA is diffrent.

```
attach(GPA)
n=5000
rand.dist=rep(NA,n)
orig=mean(c_f)-mean(c_so)-mean(c_j)-mean(c_se)
for(i in 1:n){
  sample_class=sample(ind)
  sample_fresh=values[sample_class=="Freshman"]
  sample_sophomore=values[sample_class=="Sophomore"]
  sample_junior=values[sample_class=="Junior"]
  sample_senior=values[sample_class=="Senior"]
  rand.dist[i]=mean(sample_fresh)-mean(sample_sophomore)-mean(sample_junior)-mean(sample_senior)
}

hist(rand.dist)
```




```r
pval=mean(abs(rand.dist)>abs(orig))
pval
```

```
## [1] 0.0572
```

At $\alpha = 0.5$,because the p-value 0.0596 is larger than α,we could not say that there is statistical difference among the class years.At $\alpha = 0.01$,because the p-value is less than α,we could say there is statistical difference among the class years.