ANOVA with R

Topics

I: One-Way ANOVA

II: Two-Way ANOVA

a. Single observation per cell

b. Balanced observations per cell

c. Unbalanced samples per cell

III: Three-Way ANOVA

I: One-Way ANOVA with R

1. Use data "chickwts" to apply One-Way ANOVA. (Testing $H_0: \mu_0 = \mu_1 = \cdots = \mu_k$)
2. Use Tukey's Method for Pair wise Comparisons (After H0 above is rejected, test pair-wise equality in each population means). (Testing $H_0: \mu_i = \mu_j$)

```
> data(chickwts)
> chickwts
weight     feed               14    141   linseed
1     179 horsebean           15    260   linseed
2     160 horsebean           16    203   linseed
3     136 horsebean           17    148   linseed
4     227 horsebean           18    169   linseed
5     217 horsebean           19    213   linseed
6     168 horsebean           20    257   linseed
7     108 horsebean           21    244   linseed
8     124 horsebean           22    271   linseed
9     143 horsebean           23    243   soybean
10    140 horsebean           24    230   soybean
11    309   linseed
12    229   linseed
13    181   linseed
```

=>3 columns, ID, Weight, Feed Type, 71 observations.

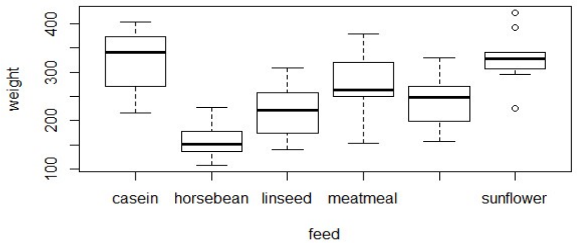
=>We test if the chickens' weights are different depending on the feed type.

```
> summary(chickwts)
    weight           feed
 Min.   :108.0   casein   :12
 1st Qu.:204.5   horsebean:10
 Median :258.0   linseed  :12
 Mean   :261.3   meatmeal :11
 3rd Qu.:323.5   soybean  :14
 Max.   :423.0   sunflower:12
```

=> There are 6 different levels of a treatment variable "feed". Each have different numbers of observations.

=> Let's take a look at the data

```
>plot(weight~feed,data=chickwts)
```



ANOVA

```
> fit<-aov(weight~feed,data=chickwts)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

=>Reorganize it as we discussed during the class...

	DF	SS	MS	F
Between Groups	K-1	SSG	MSG =SSG/K-1	MSG/MSW
Within Groups	N-K	SSW	MSW=SSQ/N-K	

Reject H0 if $\frac{\mathrm{MSG}}{\mathrm{MSW}} > \mathrm{F}(\mathrm{k}-1, \mathrm{N}-\mathrm{K}, \alpha)$

	DF	SS	MS	F
Between Groups	5	231129	46226	15.365
Within Groups	65	195556	3009	

Since F value (F(5,65,0.95)) is found as:

```
> qf(0.95,df1=5,df2=65)
[1] 2.356028
```

=> MSG/MSW > F(5,65,0.95), Conclude to reject H0.

=> Then test pair-wise comparisons.

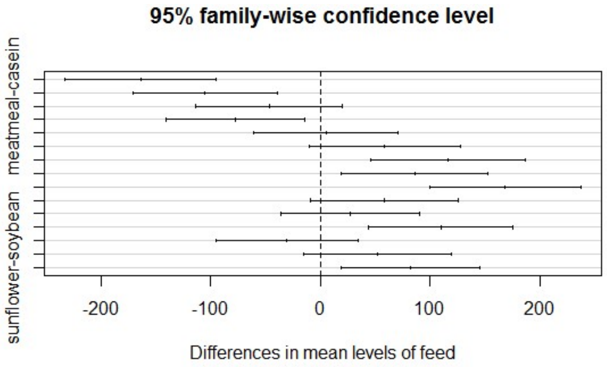
2. Tukey's Method for Pair wise Comparisons

```
> posthoc<-TukeyHSD(x=fit,'feed',conf.level=0.95,data=chickwts)
> print(posthoc)
  Tukey multiple comparisons of means
    95% family-wise confidence level

                         diff        lwr       upr     p adj
horsebean-casein   -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein     -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein     -46.674242 -113.906207  20.55772 0.3324584
soybean-casein      -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein      5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean    58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean  116.709091   46.335105 187.08308 0.0001062
soybean-horsebean    86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean 168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed     58.159091   -9.072873 125.39106 0.1276965
soybean-linseed      27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed   110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal    -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal   52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean    82.488095   19.125803 145.85039 0.0038845
```

=> The result of Tukey's method indicates that population means are **different** for the following combinations:

```
> plot(posthoc)
```

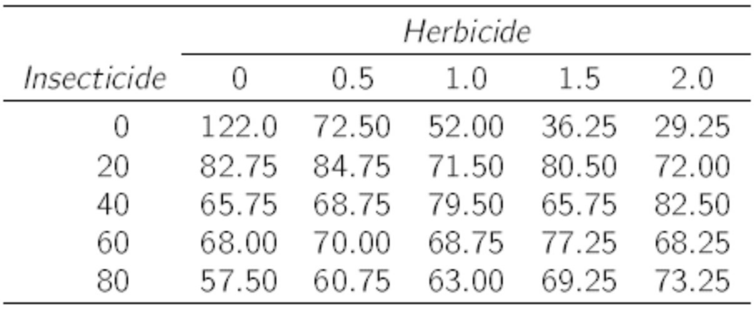
1. hoursebean-casein
2. linseed-casein
4.soybean-casein
7.meatmeal-hoursebean
8. soybean-horsebean
9. sunflower-horsebean
12. sunflower-linseed
15. sunflower-soybean



II: Two-Way ANOVA

a. Single Observation per Cell

1. Use cotton.csv data to conduct Two-way ANOVA.



	Herbicide				
Insecticide	0	0.5	1.0	1.5	2.0
0	122.0	72.50	52.00	36.25	29.25
20	82.75	84.75	71.50	80.50	72.00
40	65.75	68.75	79.50	65.75	82.50
60	68.00	70.00	68.75	77.25	68.25
80	57.50	60.75	63.00	69.25	73.25

This is 5 x 5 factorial design. Treatment 1 has 5 levels and Treatment 2 also has 5 levels. There are total 25 treatment combinations, and in this example, only single observation is observed for each treatment. The produced weight is the experimental unit, and it is the output of each combination of treatments. When you enter the data in ___ . csv format, you enter the data as follow (1st column: ID, 2nd column, insecticide levels, 3rd column, herbicide levels for each level of insecticide, and the last column output.)

Import data:

```
> cotton <- read.csv("C:/Shihomi/Shihomi_Office/Courses_HU/2015_2016_Guz/ECO663/LectureNotes_2015_2016/cotton.csv")
> cotton
  insecticide herbicide weight
1         0     0.0 122.00
2         0     0.5  72.50
3         0     1.0  52.00
4         0     1.5  36.25
5         0     2.0  29.25
6        20      0.0  82.75
7        20      0.5  84.75
8        20      1.0  71.50
9        20      1.5  80.50
10       20      2.0  72.00
11       40      0.0  65.75
12       40      0.5  68.75
13       40      1.0  79.50
14       40      1.5  65.75
15       40      2.0  82.50
```

```
16      60     0.0  68.00
17      60     0.5  70.00
18      60     1.0  68.75
19      60     1.5  77.25
20      60     2.0  68.25
21      80     0.0  57.50
22      80     0.5  60.75
23      80     1.0  63.00
24      80     1.5  69.25
25      80     2.0  73.25
```

```
> table(cotton$insecticide,cotton$herbicide)

     0  0.5 1  1.5 2
 0   1  1   1  1  1
 20  1  1   1  1  1
 40  1  1   1  1  1
 60  1  1   1  1  1
 80  1  1   1  1  1
```

Define insecticide and herbicide variables as factors.

```
> cotton$insecticide<-factor(cotton$insecticide)
> cotton$herbicide<-factor(cotton$herbicide)

> fit<-aov(weight~insecticide+herbicide,data=cotton)
> summary(fit)
             Df Sum Sq Mean Sq F value Pr(>F)
insecticide   4   799    199.8   0.583  0.680
herbicide     4   687    171.7   0.501  0.736
Residuals    16  5489   343.1
```

Two Way ANOVA table

	DF	SS	MS	F
Between Groups	K-1	SSG	MSG=SSG/K-1	MSG/MSE
Between Blocks	H-1	SSB	MSB=SSB/H01	MSB/MSE
Error	(K-1)(H-1)	SSE	MSE = SSE/(K-1)(H-1)	

Decision Rule 1: Reject Ho: K population means are all the same (treatments)

if MSG/MSE > F[(K-1), (K-1)(H-1),α]

Decision Rule 2: Reject Ho: H population means are all the same (blocks)

if MSB/MSE > F[(H-1), (K-1)(H-1), α]

The above result indicates:

	DF	SS	MS	F
Between Groups	4	799	199.8	0.583
Between Blocks	4	687	171.7	0.501
Error	16	5489	343.1	

where F[4, 16,0.95]=3.0 > MSG/MSE or MSB/MSE.

```
> qf(0.95,df1=4,df2=16)
[1] 3.006917
```

=> Fail to reject treatment effects as well as block effects.

b. Balanced Observations per cell

(1 treatment with 3 levels, 1 blocking factor with 2 levels, 10 observations for each combination.)

Data: ToothGrowth

The effect of Vitamin C on Tooth Growth in Guinea Pig.

Variables: Tooth Length (len) = Experimental Unit

Supplement Type (supp) = (VC = ascorbic acid or OJ = orange juice)

Dose in milligrams (dose) = (0.5, 1 or 2 mg)

```
> summary(ToothGrowth)
     len        supp         dose
 Min.   : 4.20   OJ:30   Min.   :0.500
 1st Qu.:13.07   VC:30   1st Qu.:0.500
 Median :19.25           Median :1.000
 Mean   :18.81           Mean   :1.167
 3rd Qu.:25.27           3rd Qu.:2.000
 Max.   :33.90           Max.   :2.000
```

```
> ToothGrowth
   len supp dose
1   4.2   VC  0.5
2  11.5   VC  0.5
3   7.3   VC  0.5
4   5.8   VC  0.5
5   6.4   VC  0.5
6  10.0   VC  0.5
7  11.2   VC  0.5
8  11.2   VC  0.5
9   5.2   VC  0.5
10  7.0   VC  0.5
11 16.5   VC  1.0
12 16.5   VC  1.0
13 15.2   VC  1.0
14 17.3   VC  1.0
15 22.5   VC  1.0
```

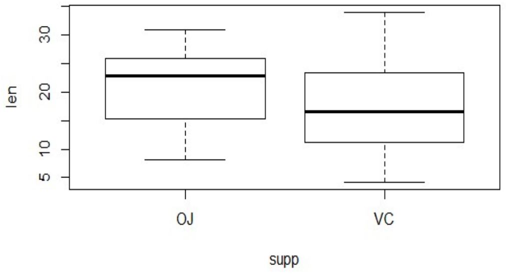
.... 60 observations

```
> table(ToothGrowth$supp, ToothGrowth$dose)
      0.5  1  2
  OJ  10 10 10
  VC  10 10 10
```

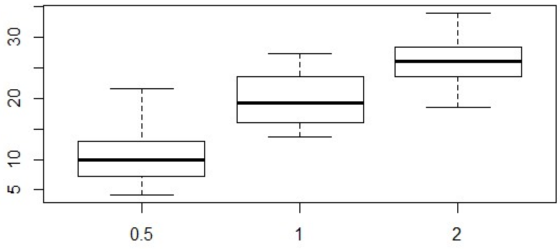
=> 10 observations per each cell.

Let's see the response versus each of the factors.

```
> boxplot(len~supp, data=ToothGrowth)
```



```
>boxplot(len~dose,data=ToothGrowth)
```



Data Preparation

Define "dose" variable as a factor. (R will recognize it as a factor, not numerical values)

```
> ToothGrowth$dose=factor(ToothGrowth$dose)
> contrasts(ToothGrowth$dose)  (Dummy-coded)
      1 2
0.5   0 0
1     1 0
2     0 1
```

Conduct a "Two Way Factorial Design" analysis.

```
> fit<-aov(len~dose+supp+dose:supp,data=ToothGrowth)   ##(aov(len~dose*supp,data=ToothGrowth)) is also equivalent.
> summary(fit)
            Df Sum Sq  Mean Sq F value   Pr(>F)
dose         2  2426.4  1213.2   92.000  < 2e-16 ***
supp         1   205.4   205.4   15.572 0.000231 ***
dose:supp    2   108.3    54.2    4.107 0.021860 *
Residuals   54   712.1    13.2
---
```

	DF	SS	MS	F
Between Groups	K-1	SSG	MSG=SSG/K-1	MSG/MSE
Between Blocks	H-1	SSB	MSB=SSB/H01	MSB/MSE
Interaction	(K-1)(H-1)	SSI	MSI = SSI/(K-1)(H-1)	MSI/MSE
Error	KH(L-1)	SSE	MSE=SSE/KH(L-1)	

	DF	SS	MS	F
Between Groups	2	2426.4	1213.2	92.00
Between Blocks	1	205.4	205.4	15.572
Interaction	2	108.3	54.2	4.107
Error	54	712.1	13.2	

```
> qf(0.95,df1=2,df2=54)
[1] 3.168246
> qf(0.95,df1=1,df2=54)
[1] 4.019541
```

Decision Rules:

A: Reject H0 (K population group (treatment) means are all the same if MSG/MSE > F(K-1,KH(L-1),α)

B: Reject H0 (H population block means are all the same) if MSB/MSE>F(H-1,KH(L-1),α)

C: Reject H0 (There is no interaction effect between Group (Treatment) and Block)
if MSI/MSE > F((K-1)(H-1), KH(L-1), α).

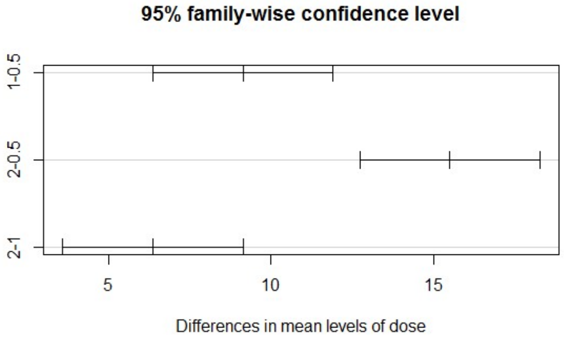
Therefore, All null hypotheses are rejected.

Tukey's method for pair-wise comparison:

```
> posthoc1<-TukeyHSD(x=fit,'dose',conf.level=0.95,data=ToothGrowth)
> print(posthoc1)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = len ~ supp * dose, data = ToothGrowth)
$dose
          diff     lwr         upr          p adj
1-0.5     9.130   6.362488   11.897512   0.0e+00
2-0.5    15.495  12.727488   18.262512   0.0e+00
2-1       6.365   3.597488    9.132512    2.7e-06 (= 0.0000027)
```

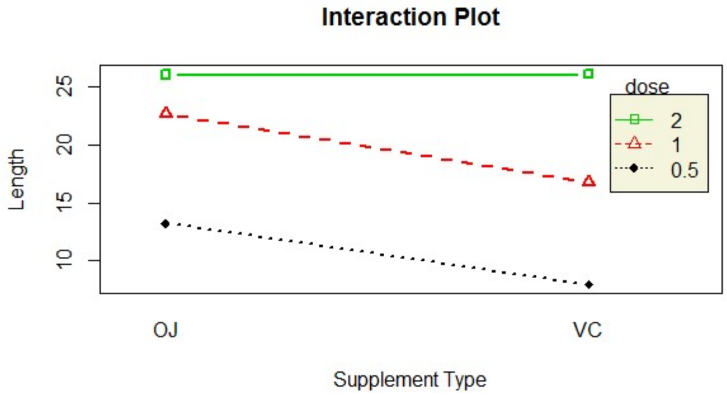
=> The effects between 1 and 0.5mg are statistically significantly different.
=> The effects between 2 and 0.5mg are statistically significantly different.
=> The effect between 2 and 1mg are also statistically significantly different.

```
>plot(posthoc1)
```



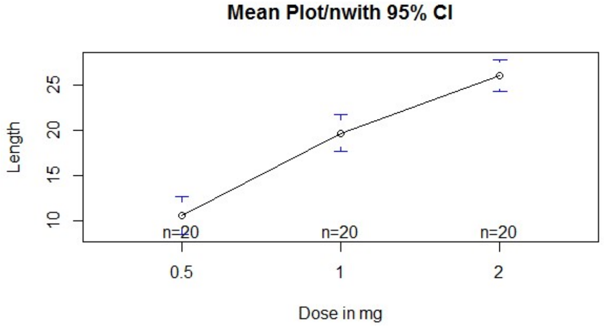
Visualization of the Result

```
> attach(ToothGrowth)
> dose <- factor(dose)
> supp <- factor(supp)
> interaction.plot(supp, dose, len, type="b", col=c(1:3),
+             leg.bty="o", leg.bg="beige", lwd=2, pch=c(18,24,22),
+             xlab="Supplement Type",
+             ylab="Length",
+             main="Interaction Plot" )
```



After installing gplots package to R, you can plot means with error bars.

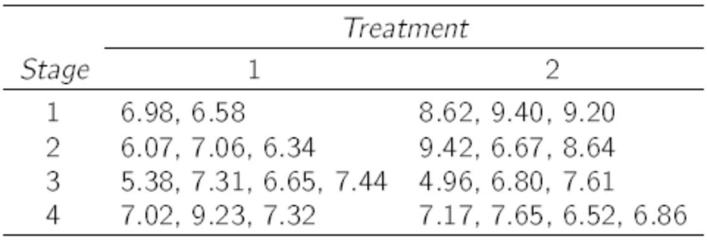
```
>library(gplots)
> dose<-factor(dose)
> plotmeans(len~dose,xlab="Dose in mg", ylab="Length", main="Mean Plot/nwith 95% CI")
```



c. Unbalanced Observations per cell

Data: adrenal.csv

An experiment to measure the effect of ACTH on the adrenal glands of rats. Two different treatments were applied to glands at four different stages of development. The response is steroid production per 100 mg of gland per hour. Four replicates per cell were planned, but some lab results were invalid, so the final dataset is unbalanced.



	Treatment	
Stage	1	2
1	6.98, 6.58	8.62, 9.40, 9.20
2	6.07, 7.06, 6.34	9.42, 6.67, 8.64
3	5.38, 7.31, 6.65, 7.44	4.96, 6.80, 7.61
4	7.02, 9.23, 7.32	7.17, 7.65, 6.52, 6.86

```
> adrenal <- read.csv("C:/Shihomi/Shihomi_Office/Courses_HU/2015_2016_Guz/ECO663/LectureNotes_2015_2016/adrenal.csv")
>   View(adrenal)
> attach(adrenal)
  stage treatment steroid
1     1         1    6.98
2     1         2    6.58
3     1         2    8.62
4     1         2    9.40
5     1         2    9.20
6     2         1    6.07
7     2         1    7.16
8     2         1    6.34
9     2         2    9.42
10    2         2    6.67
11    2         2    8.64
12    3         1    5.38
13    3         1    7.31
14    3         1    6.65
15    3         1    7.44
16    3         2    4.96
17    3         2    6.80
18    3         2    7.61
19    4         1    7.02
20    4         1    9.23
21    4         1    7.32
22    4         2    7.17
23    4         2    7.65
24    4         2    6.52
25    4         2    6.86
```

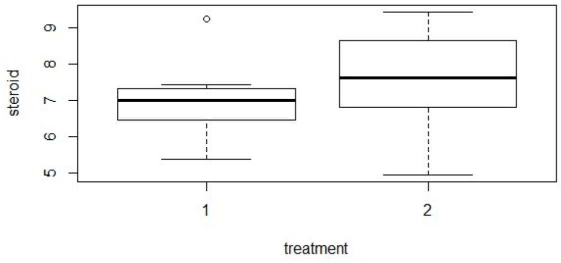
```
> table(adrenal$stage,adrenal$treatment)
     1  2
  1  2  3
  2  3  3
  3  4  3
  4  3  4
<= Number of observations per cell.
```

	Treatment level 1	Treatment level 2
Stage 1	2	3
Stage 2	3	3
Stage 3	4	3
Stage 4	3	4

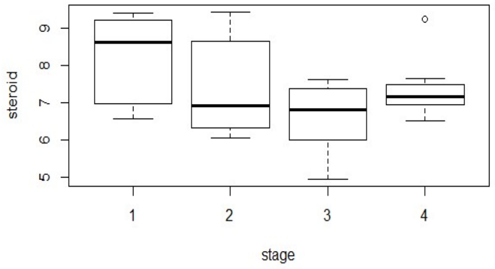
=> The numbers of observations are different for each combination of the treatment.

Convert "stage" and "treatment" variables as factors.

```
> adrenal$stage<-factor(adrenal$stage)
> adrenal$treatment<-factor(adrenal$treatment)
> plot(steroid~treatment,data=adrenal)
```



```
> plot(steroid~stage,data=adrenal)
```



```
> fit4<-aov(steroid~stage*treatment,data=adrenal)
> summary(fit4)
                Df Sum Sq Mean Sq F value Pr(>F)  
stage            3  7.260   2.420   2.745 0.0751 .
treatment        1  2.045   2.045   2.320 0.1461  
stage:treatment  3  9.916   3.305   3.749 0.0311 *
Residuals       17 14.988   0.882                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

<=fit4 line is equivalent to fit4<-aov(steroid~stage+treatment+stage:treatment, data=adrenal)

=> Steroid levels are different between stages. (10% significance level) (H0 is rejected at 10% level)
=> Steroid levels are not statistically different between treatment levels. (H0 cannot be rejected)
=> Steroid levels are different depending on the combination of stage and treatment. (H0 rejected at 5% level).

Let's now run a linear estimation model to see the differences/similarities in the implications.
# Changing dummy-coding to effect-coding of treatment and stage variables.

```
> contrasts(adrenal$stage) <- contr.sum
> contrasts(adrenal$treatment) <- contr.sum
```

#Fit a linear model

```
> result<-lm(steroid~stage*treatment, data=adrenal)
> summary(result)

Call:
lm(formula = steroid ~ stage * treatment, data = adrenal)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5733 -0.4533  0.1200  0.6000  1.3733 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        7.33479    0.19166  38.269   <2e-16 ***
stage1             0.59188    0.35857   1.651   0.1172    
stage2             0.04854    0.33197   0.146   0.8855    
stage3            -0.75896    0.31784  -2.388   0.0288 *  
treatment1        -0.37104    0.19166  -1.936   0.0697 .  
stage1:treatment1 -0.77562    0.35857  -2.163   0.0451 *  
stage2:treatment1 -0.48896    0.33197  -1.473   0.1591    
stage3:treatment1  0.49021    0.31784   1.542   0.1414    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.939 on 17 degrees of freedom
Multiple R-squared:  0.5619,	Adjusted R-squared:  0.3815 
F-statistic: 3.115 on 7 and 17 DF,  p-value: 0.02631
```

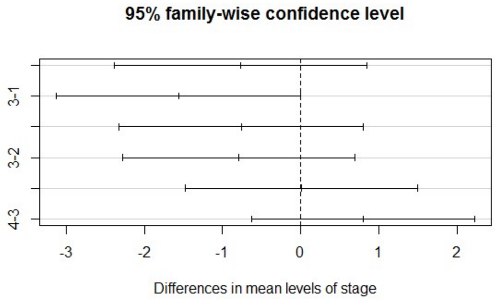
=> The result indicates that

1. stage3 effect is statistically significantly lower than the effect of stage 4.
2. treatment effect 1 is lower than treatment effect 2 at 10% significance level.
3. interaction of "stage1 + treatment 1" effect is statistically significantly lower than "stage1 + treatment 2" effect at 5% significance level. But there is no other interaction which is significant.

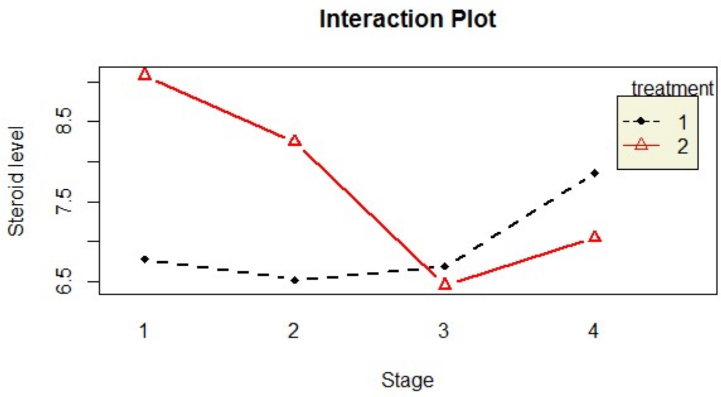
```
> posthoc=TukeyHSD(x=fit,'stage',conf.level=0.95,data=adrenal)
> print(posthoc)
 Tukey multiple comparisons of means
   95% family-wise confidence level

Fit: aov(formula = steroid ~ stage * treatment, data = adrenal)

$stage
             diff          lwr         upr              p adj
2-1     -0.77266667 -2.3888588  0.8435254621   0.5402666
3-1     -1.56314286 -3.1259805 -0.0003051792   0.0499458
4-1     -0.76028571 -2.3231234  0.8025519636   0.5262257
3-2    -0.79047619 -2.2754005  0.6944480710    0.4516831
4-2     0.01238095 -1.4725433  1.4973052139    0.9999950
4-3     0.80285714 -0.6238119  2.2295262261    0.4048957
```

=> Therefore, there is a statistically different effect between stages 1 and 3, but not others.

```
> plot(posthoc)
```



```
> interaction.plot(stage, treatment,steroid, type="b", col=c(1:3),leg.bty="o", leg.bg="beige", lwd=2, pch=c(18,24,22),xlab="Stage", ylab="S
teroid level",main="Interaction Plot")
>
```



III: Three-Way ANOVA

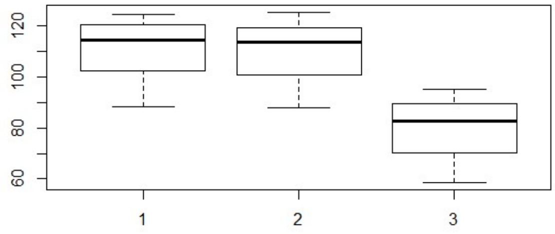
Data: marketing.csv

Factors :
A: fee (1=high, 2=average,3=low)
B: Scope (1=all work performed in house, 2=some work subcontracted out)
C: supervision (1=local supervisors, 2=traveling supervisors only).

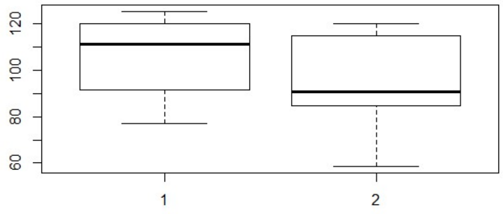
3x2x2 factorial design.
Experimental unit: Quality of work done by marketing research agencies.

```
> marketing <- read.csv("C:/Shihomi/Shihomi_Office/Courses_HU/2015_2016_Guz/ECO663/LectureNotes_2015_2016/marketing.csv")
> attach(marketing)
> table(marketing$scope,marketing$fee)        > table(marketing$supervision, marketing$fee)
   1 2 3                                         1 2 3
 1 8 8 8                                       1 8 8 8
 2 8 8 8                                       2 8 8 8
> table(marketing$scope,marketing$supervision,marketing$fee)
, ,  = 1                            , ,  = 2                            , ,  = 3
   1 2                                 1 2                                 1 2
 1 4 4                               1 4 4                               1 4 4
 2 4 4                               2 4 4                               2 4 4
```

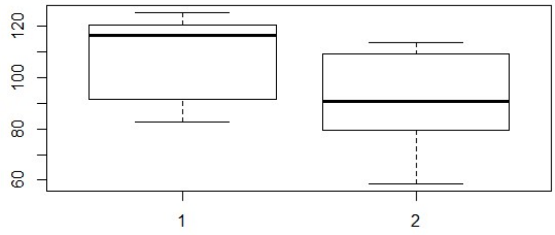
```
>boxplot(quality~fee,data=marketing)
```



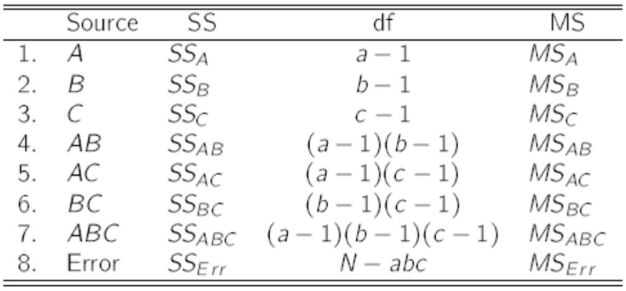
```
> boxplot(quality~scope,data=marketing)
```



```
> boxplot(quality~supervision,data=marketing)
```



```
> marketing$fee=factor(marketing$fee)
> marketing$scope=factor(marketing$scope)
> marketing$supervision=factor(marketing$supervision)
> fit=aov(quality~fee*scope*supervision,data=marketing)
> summary(fit)
                      Df Sum Sq Mean Sq F value  Pr(>F)
fee                    2  10044    5022 679.336 < 2e-16 ***
scope                  1   1834    1834 248.079 < 2e-16 ***
supervision            1   3832    3832 518.403 < 2e-16 ***
fee:scope              2      2       1   0.108   0.898
fee:supervision        2      1       0   0.053   0.948
scope:supervision      1    575     575  77.749 1.6e-10 ***
fee:scope:supervision  2      4       2   0.267   0.767
Residuals             36    266       7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```



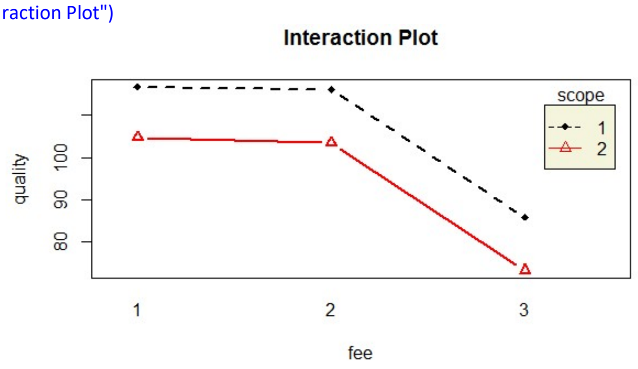
	Source	SS	df	MS
1.	A	SS_A	$a-1$	MS_A
2.	B	SS_B	$b-1$	MS_B
3.	C	SS_C	$c-1$	MS_C
4.	AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}
5.	AC	SS_{AC}	$(a-1)(c-1)$	MS_{AC}
6.	BC	SS_{BC}	$(b-1)(c-1)$	MS_{BC}
7.	ABC	SS_{ABC}	$(a-1)(b-1)(c-1)$	MS_{ABC}
8.	Error	SS_{Err}	$N-abc$	MS_{Err}

	DF	SS	MS	F	
Fee	2	10044	5022	679.336***	①
Scope	1	1834	1834	248.079***	②
Supervision	1	3832	3832	518.403***	③
Fee:Scope	2	2	0.108	0.898	④
Fee:Supervision	2	1	0	0.948	⑤
Scope:Supervision	1	575	575	77.749***	⑥
Fee:Scope:Supervision	2	4	2	0.267	⑦
Residuals	36	266	7		

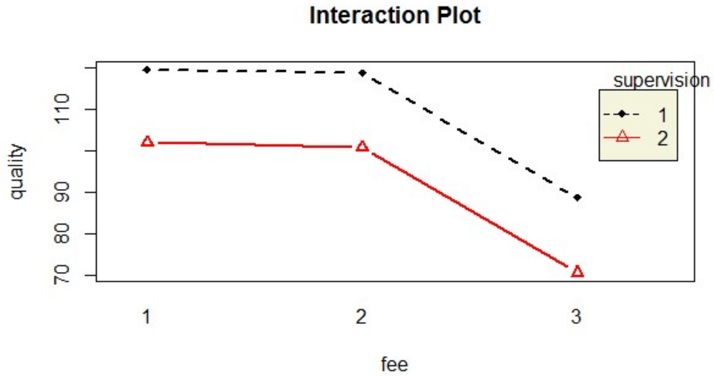
The result indicates:

1. Each variable has a strong main effect.
2. A two-way interaction exists between factors B (scope) and C (supervision).
3. All the other interactions are very small and negligible.

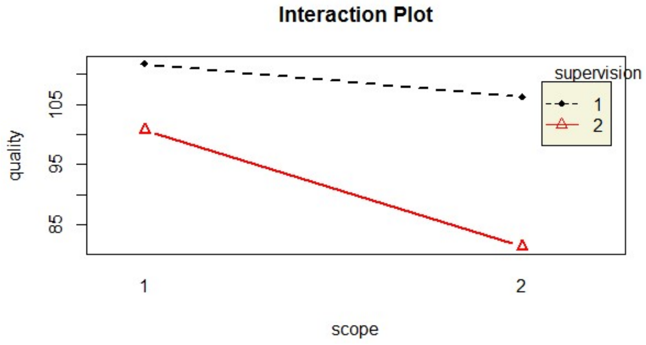
```
> interaction.plot(fee,scope,quality,type="b",col=c(1:3),leg.bty="o",leg.bg="beige",lwd=2,pch=c(18,24,22),xlab="fee",ylab="quality",main="Interaction Plot")
```



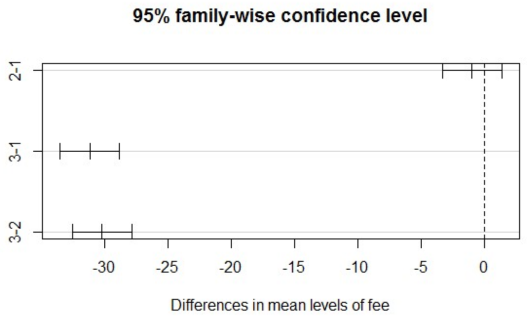
```
>interaction.plot(fee,supervision,quality,type="b",col=c(1:3),leg.bty="o",leg.bg="beige",lwd=2,pch=c(18,24,22),xlab="fee",ylab="quality",main=
"Interaction Plot")
```



```
>interaction.plot(scope,supervision,quality,type="b",col=c(1:3),leg.bty="o",leg.bg="beige",lwd=2,pch=c(18,24,22),xlab="scope",ylab="quality",
main="Interaction Plot")
```



```
> posthoc=TukeyHSD(x=fit,'fee',conf.level=0.95, data=marketing)
> print(posthoc)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = quality ~ fee * scope * supervision, data = marketing)

$fee
         diff        lwr        upr     p adj
2-1  -0.96250  -3.312191   1.387191 0.5808587
3-1 -31.15625 -33.505941 -28.806559 0.0000000
3-2 -30.19375 -32.543441 -27.844059 0.0000000

> plot(posthoc)
```


95% family-wise confidence level
Differences in mean levels of fee