STAT 5310 (Winter 2021) - Assignment 3

Robin, Teotia, TRU ID: T00671961

Due: 11:00 PM, Mar. 19, 2021

Due date: Electronic submission on Moodle Learning Portal course page by Friday, Mar. 19, 2021 at 22:00. NB: e-mail submissions will NOT be accepted unless you have some special situation to accommodate.

If you work with other students on this assignment then:

- indicate the names of the students on your solutions.
- your solutions must be written up independently (i.e. your solution should not be the same as another students solutions).

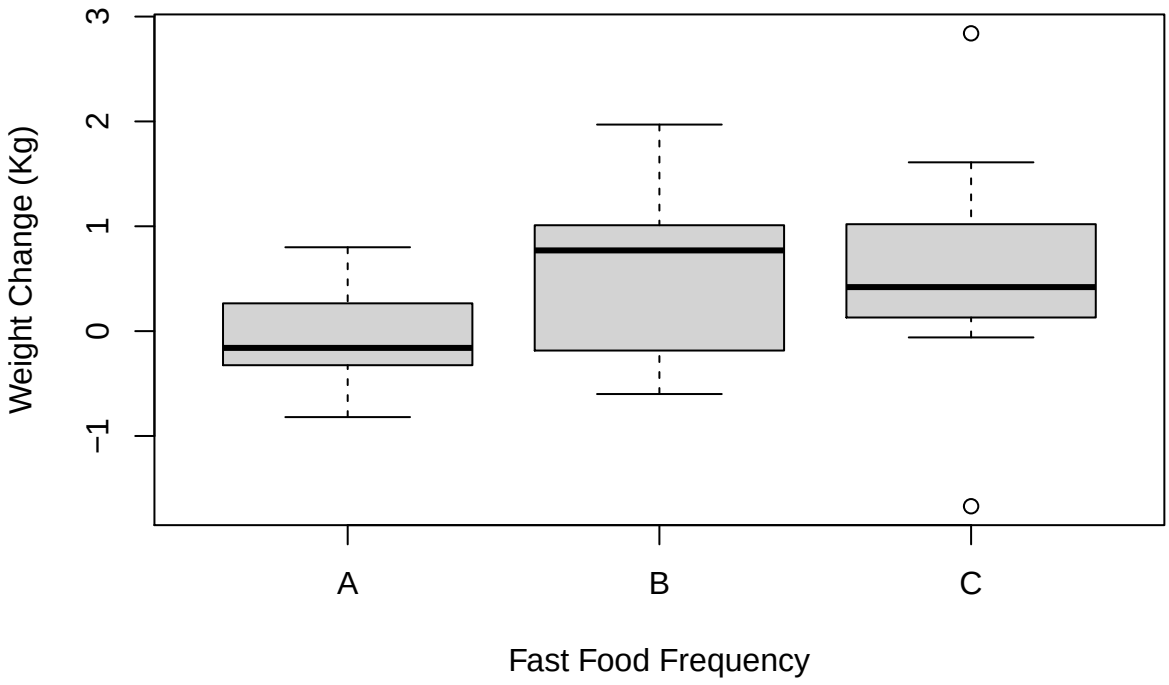
Q1 [20 pts]

A study at TRU recruited twenty-one students to complete a thirty minute survey on diet and eating habits during the academic year. Students were paid $10 to complete the survey. The data below shows their weight gain from September to April classified by the frequency that students ate fast food. In group A students reported never eating fast food; students in group B reported eating fast food twice per month; and students in group C reported eating fast food four times per month.

Information	A	B	C
	0.52	1.14	-0.06
	-0.82	0.10	0.32
	-0.23	-0.60	2.84
	-0.42	1.97	-1.67
	0.01	-0.47	0.42
	-0.16	0.77	0.43
	0.80	0.88	1.61
Treatment Average	-0.04	0.54	0.56
Treatment SD	0.55	0.92	1.40

The researchers analyzed the data using R.

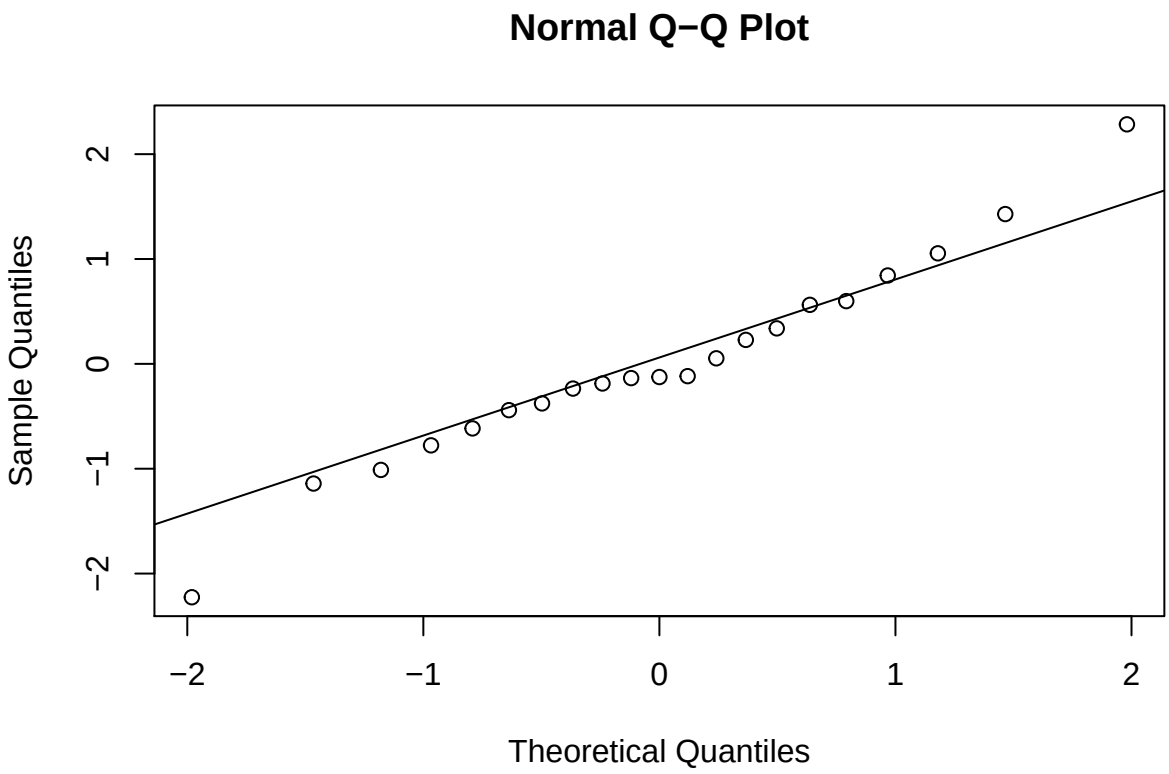
```
data <- read.csv("surveydata.csv")
grp <- as.factor(data$grp)
boxplot(y~grp,data = data,ylab="Weight Change (Kg)",
        xlab="Fast Food Frequency")
```



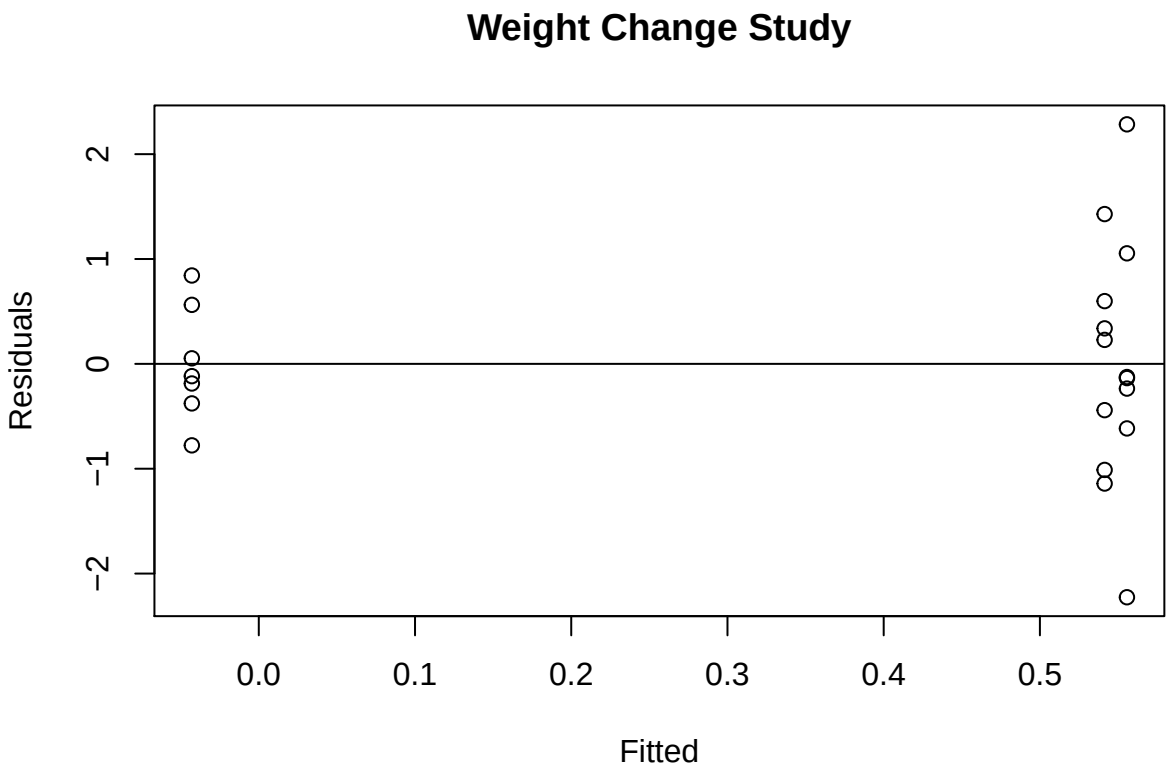

```
aovdata <- aov(y~grp,data=data)
summary(aovdata)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## grp          2  1.633  0.8165   0.787   0.47
## Residuals   18 18.664  1.0369
```

```
qqnorm(aovdata$residuals);qqline(aovdata$residuals)
```




```
plot(aovdata$fitted.values, aovdata$residuals,
     ylab="Residuals",
     xlab="Fitted",
     main="Weight Change Study")
abline(h=0)
```



Answer the following questions. (20 marks)

(a) [4 pts] Is this study an experiment or observational study? What is the treatment? Briefly explain.

Ans: It is an observational studies as the treatments are not assigned ny the resarcher.

(b) [6 pts] Would it have been feasible for the researcher to randomize students to the treatments? What randomization scheme (assigning the subjects to the treatments) could the researcher use to accomplish the randomization? Briefly explain.

Ans: No, it is not feasible for the researcher to randomise students to the treatments(that is to eat fast food at acertain number of times per months).This study could be randomised, by appling the random permutation after assinging the numbers to the subjects from 1 to 21.Here we will make 3 groups,the first 7 would be assigned to group A and next 7 to the group B and so on and so forth.The way to do this is:sample(1:21).

(c) [6 pts] What is the ANOVA identity for this data? Is it necessary that the data collected in each treatment group follow a normal distribution for the ANOVA identity to be true? Briefly explain.

Ans:The ANOVA identity is SSTotal=SSGroup +SSResid. From the above data: 20.297= 1.633+18.664.The anova identity is true regardless of the distribution within each group, as it follows the Pythagoras' theorem in higher dimensions.

(d) [4 pts] Two plots are produced in the analysis, which assumption are they checking, respectively?

Ans: plot1:checking the assumptions for normality of the residuals i.e errors are normally distributed "normality okay", plot2: checking the assumptions of constant variance,the points are falling randomly on both sides of 0, with no recognizable patterns in the points, it means satified.

Q2 [20 pts]

A clinical trial was conducted where patients were randomized to four different treatments. The data is available in the file `q2data.csv`. The outcome is a continuous response y_{ij} the response for the ith subject in the jth treatment group. There are three new treatments in this study and one control treatment. **The control treatment is the third treatment ($j = 3$)**. The main objective of the study is to compare the three new treatments to the control treatment.

NB: The file can be read into R and put into a data.frame using the command

```
q2data <- read.csv("q2data.csv")
```

In this question use the 5% significance level.

(a) [5 pts] What are the averages and standard deviations of each treatment? Plot the distributions of the four treatment groups. Do the distributions look similar or different? (Hand in your R code and output)

```
q2data <- data.frame(read.csv("q2data.csv"))
attach(q2data)
head(q2data)
```

```
##   X          y trt
## 1 1  0.2860652   1
## 2 2  1.8354438   1
## 3 3  1.6397704   1
## 4 4  2.8415955   1
## 5 5 -0.2231735   1
## 6 6  1.8625634   1
```

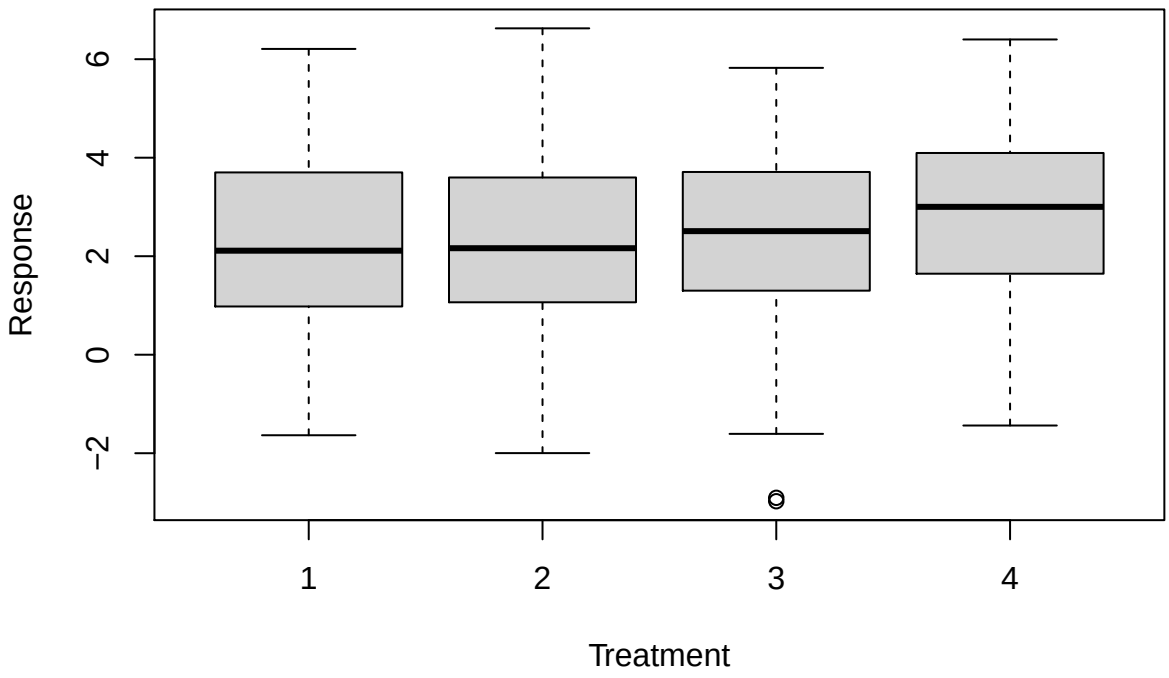
```
#Average for each treatment
sapply(split(y,trt),mean)
```

```
##        1        2        3        4
## 2.206182 2.290470 2.320007 2.855205
```

```
#Standard Deviation for each treatment
sapply(split(y,trt),sd)
```

```
##        1        2        3        4
## 1.799560 1.774576 1.812561 1.763111
```

```
#Distribution of the four treatment groups
boxplot(y~trt,ylab = "Response",xlab = "Treatment")
```



Ans: The distribution looks different as there is difference in the plotted data like the medians,quantile and extremes values.

(b) [5 pts] Use linear regression to calculate the ANOVA table. What do you conclude from the ANOVA table? (NB: when using linear regression to calculate the effects the treatment variable should be specified as a factor `as.factor(trt)`.) (Hand in your R code and output)

```
# write your R code here
#making levels
treatment<-as.factor(trt)
#linear regression
lm(y~treatment)
```

```
##
## Call:
## lm(formula = y ~ treatment)
##
## Coefficients:
## (Intercept)   treatment2   treatment3   treatment4
##     2.20618      0.08429      0.11382      0.64902
```

```
#Anova through linear regression
anova(lm(y~treatment))
```

```
## Analysis of Variance Table
```

```
##
## Response: y
##            Df  Sum Sq Mean Sq F value  Pr(>F)
## treatment   3   26.29  8.7639  2.7429 0.04294 *
## Residuals 391 1249.31  3.1952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ans: Ans:a-1=3,N-a=391, Null Hypothesis(H0): $\mu 1 = \mu 2 = \mu 3 = \mu 4$,Alternative Hypotheses(Ha):At least two treatment has different mean result.As the P value is 0.04294<0.05, hence we reject the null hypothesis and conclude that there exist at lest two treatments having different mean results.

(c) [5 pts] Use the model you obtained in part (b) to obtain the appropriate parameter estimates using the treatment contrast (dummy coding) to answer the main objective. In R this can be done using the `contr.treatment()` function. Define the underlying statistical model in terms of dummy variables. Explicitly state the dummy variables. Interpret the parameter estimates. Verify the parameter estimates using the table of means that you obtained in part (a). (Hand in your R code and output)

```
# write your R code here
contrasts(treatment) <- contr.treatment(4,base = 3)
contrasts(treatment)
```

```
##   1 2 4
## 1 1 0 0
## 2 0 1 0
## 3 0 0 0
## 4 0 0 1
```

```
summary(lm(y~treatment))
```

```
##
## Call:
## lm(formula = y ~ treatment)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.295 -1.180 -0.048  1.391  4.336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.32001    0.18056  12.849   <2e-16 ***
## treatment1  -0.11382    0.25408  -0.448   0.6544
## treatment2  -0.02954    0.25668  -0.115   0.9084
## treatment4   0.53520    0.25345   2.112   0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.788 on 391 degrees of freedom
## Multiple R-squared:  0.02061,    Adjusted R-squared:  0.0131
## F-statistic: 2.743 on 3 and 391 DF,  p-value: 0.04294
```

X1 ={1 for treatment 1,else 0}.
X2 ={1 for treatment 2,else 0}.

X4 ={1 for treatment 4,else 0}.
Ans: There are 4 treatment so they only need 3 dummy variables.The 3 Dummy variables are X1,X2 and X4 and here the reference level is the third treatment that is control treatment..
intercept = mean of 3 level=2.32001.
β_i = mean of i level - mean of 3 level.
β_1 = mean of 1 level - mean of 3 level=-0.11382l.
This implies mean of treatment_1=2.32001-0.11382= 2.20619.
β_2 = mean of 3 level - mean of 3 level=-0.02954.
This implies mean of treatment_2=2.32001-0.02954 =2.29047.
β_4 = mean of 4 level - mean of 3 level=0.53520.
This implies mean of treatment_4=2.32001+0.53520=2.85521.
The mean we got here of all the treatment level is exactly the same as we got in Part (a) of the question, hence verified.

(d) [5 pts] Which pairs of treatments have a statistically significant difference? Do your results change if you adjust for multiple comparisons using either the Bonferroni or Tukey method? Compare all pairs of treatment means using no adjustement, Bonferroni, and Tukey. If the unadjusted, Bonferroni, and Tukey lead to different conclusions then explain why these methods give different results. Does it make sense to consider all pairs of treatment means given the main objective of this study? (Hand in your R code and output)

```
# Write your code here
#unadjusted p-values
pairwise.t.test(y,treatment,p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and treatment
##
##   1     2     3
## 2 0.742 -     -
## 3 0.654 0.908 -
## 4 0.010 0.027 0.035
##
## P value adjustment method: none
```

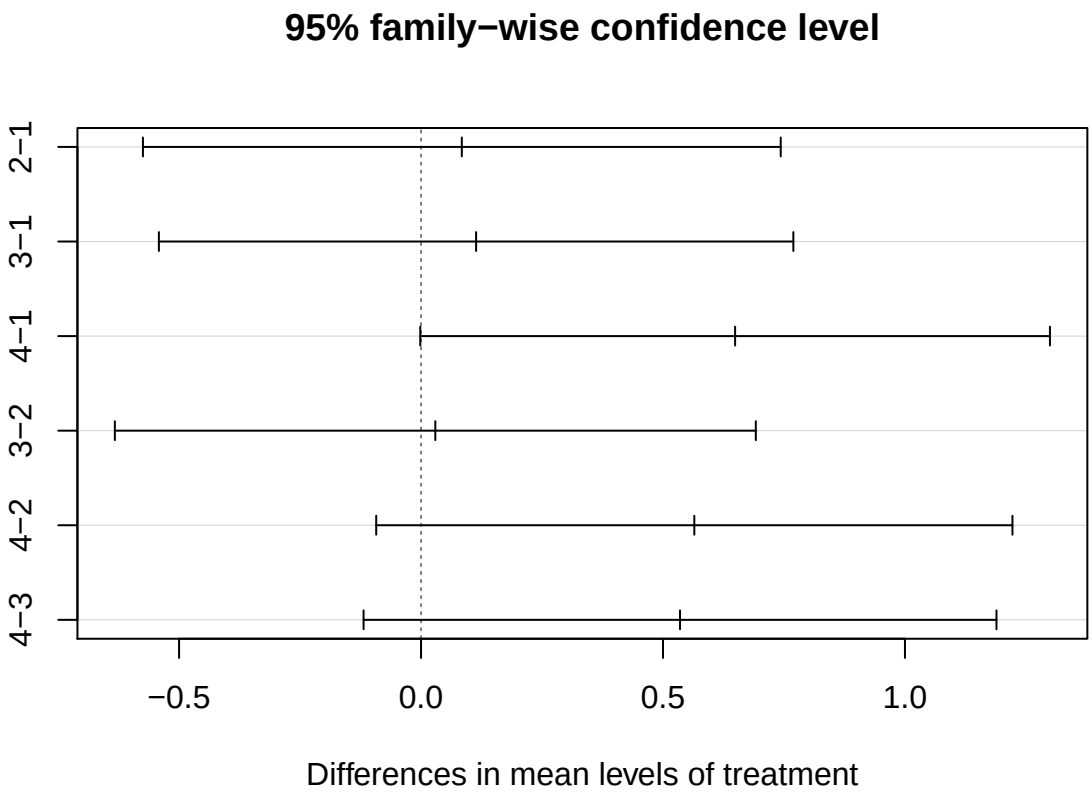
```
#Bonferroni adjusted p-values
pairwise.t.test(y,treatment,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and treatment
##
##   1     2     3
## 2 1.000 -     -
## 3 1.000 1.000 -
## 4 0.063 0.163 0.212
##
## P value adjustment method: bonferroni
```

```r
#Tukey adjusted p-value
round(TukeyHSD(aov(y~treatment,data=q2data))$treatment,2)
```

```
##     diff   lwr  upr p adj
## 2-1 0.08 -0.57 0.74  0.99
## 3-1 0.11 -0.54 0.77  0.97
## 4-1 0.65  0.00 1.30  0.05
## 3-2 0.03 -0.63 0.69  1.00
## 4-2 0.56 -0.09 1.22  0.12
## 4-3 0.54 -0.12 1.19  0.15
```

```r
plot(TukeyHSD(aov(y~treatment,data=q2data)))
```



Ans: No pair of tretment has statistically significant differnece. Here in unadjusted, the significance level is α/c =0.05/6=0.0083.Hence all the received Pi values is greater than the α/c =(0.0083).Thus we fails to reject the null hypothesis, and hence there is no pair with statistically significant diffrence.My results does not change when I adjust for multiple comparisons using the Bonferroni, with α=0.05, I got no pair with statistically significant differnce. Also I get the same results as Bonferroni while appling Tukey method that is I got no pair with statiscally differnce.Hence, all the 3 methods giving the same results.No, it doesn't make sense to consider all pairs of treatment means given the main objective of this study.

Q3 [11 pts]

A data set from Oehlert (2000), *A First Course in Design and Analysis of Experiments*, New York: W. H. Freeman. In this dataset we have five Cycad plants, three fronds per plant.

- Treatment: water (control), fungal spores, and horticultural oil.
- 5 infested cycads, 3 branches are randomly chosen on each cycad, and 2 patches (3 cm $\times$ 3 cm) are marked on each branch
- 3 branches on each cycad are randomly assigned to the 3 treatments
- Response is the average change in number of mealybugs after 3 days in two 3 cm by 3 cm patches per frond.

In this question use the 5% significance level.

The columns in the data set are treatment, plan and average change.

(a) [1 pts] What is the name of this design?

Ans: The name of the design is randomised block design

(b) [4 pts] Find the ANOVA table for a randomized block design with one replication per block-treatment combination. To get such table, you should first find the response average in each block-treatment combination, and find the corresponding treatment and blocking variable. From this ANOVA table, what conclusion do you have for equal mean testing for the treatment variable at 5%? (Show your R code and the ANOVA table output)

```
# Put your R  code here
q3data_given <- read.csv("cycad-data.csv")
q3data <- aggregate(change ~ trt + plant, data=q3data_given, FUN=mean)
attach(q3data)
```

```
## The following object is masked from q2data:
##
##     trt
```

```
head(q3data)
```

```
##   trt plant change
## 1   1     1   -7.5
## 2   2     1    1.5
## 3   3     1    7.5
## 4   1     2   11.5
## 5   2     2   19.5
## 6   3     2   32.5
```

```
#The block averages are:
block.ave <- sapply(split(q3data$change,q3data$plant),mean);block.ave
```

```
##         1         2         3         4         5
##  0.500000 21.166667  8.666667  7.500000  6.500000
```

```r
#The treatments averages are:
trt.ave <- sapply(split(q3data$change,q3data$trt),mean);trt.ave
```

```
##    1    2    3
##  4.3  5.9 16.4
```

```r
#calculating grand mean
grand.ave <- mean(q3data$change);grand.ave
```

```
## [1] 8.866667
```

```r
#SSBlock
block.devs <- block.ave-grand.ave; block.devs; sum(block.devs^2)*3
```

```
##         1         2         3         4         5
## -8.366667 12.300000 -0.200000 -1.366667 -2.366667
```

```
## [1] 686.4
```

```r
#SStreat
treatment.devs <- trt.ave-grand.ave; treatment.devs; sum(treatment.devs^2)*5
```

```
##         1         2         3
## -4.566667 -2.966667  7.533333
```

```
## [1] 432.0333
```

```r
#SST
all.devs <- q3data$change-grand.ave; sum(all.devs^2)
```

```
## [1] 1260.233
```

```r
#the error sum of squares
sum(all.devs^2)-sum(treatment.devs^2)*5-sum(block.devs^2)*3
```

```
## [1] 141.8
```

```r
anova(lm(q3data$change~as.factor(q3data$trt)+as.factor(q3data$plant),data = q3data))
```

```
## Analysis of Variance Table
##
## Response: q3data$change
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(q3data$trt)     2 432.03 216.017 12.1871 0.003729 **
## as.factor(q3data$plant)   4 686.40 171.600  9.6812 0.003708 **
## Residuals                 8 141.80  17.725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ans: H0: $\mu(trt1) = \mu(trt2) = \mu(trt3)$, H1: at least one pair of treatment mean is different, here in refernce to treatment the p-value is 0.003729< 0.05(given significance level), hence we reject the null hpthothesis which further implies there exists at least one pair of treatment whose means are different.

(c) [6 pts] For (b) analysis, dropping the blocking variable, and analyzing this data set by CRD, What conclusion do you have for the treatment testing based on this analysis at 5% level? Does this conclusion agree with what you obtained in (b)? Can you explain why the results agree or disagree?

```
# put R code here
anova(lm(change~as.factor(trt),data = q3data))
```

```
## Analysis of Variance Table
##
## Response: change
##                Df Sum Sq Mean Sq F value  Pr(>F)
## as.factor(trt)  2 432.03 216.017  3.1299 0.08056 .
## Residuals      12 828.20  69.017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ans: H0: $\mu(trt1) = \mu(trt2) = \mu(trt3)$, H1: at least one pair of treatment mean is different, here the p-value is 0.08056 $>$ 0.05(given significance level), hence we fails to reject the the null hpthothesis which further implies that treatment means are equal. It is clear that the conclusion drawn here does not agree with the solution obatined in (b).SSE increases from 141.80 to 828.20. SST is not associated with treatments or the mean, almost half is accounted for by block-to-block variation. If the experiment had been arranged on a completely randomized basis with no blocks, the error variance would have been much larger. The randomized block design increased the sensitivity of this experiment.In (b) part the SSE is 141.80, mean square of the residuals is 17.725 and Mean Square of treatment=216.017 gives F-value=216.017/17.725=12.1871(High F value corresponds low p- value). In this Part, the SSE is 828.20(incresed),mean square of the residuals is 69.017(increased,denominator) and Mean Square of treatment=216.017(numerator same) gives F-value=216.017/69.017=3.1299(comparatively low F value corresponds comparatively high p- value),And this is the reason of getting diffrent and reverse results.

Q4 [4 pts] What is the name of the design in the follow scenarios

(a) [1 pt] Fifteen judges rated two randomly allocated brands of beers, A and B, according to taste (scale: 1 to 10).

Ans:Unpaired Randomization Distribution Design

(b) [1 pt] Twenty judges each rated two brands of beers, A and B, according to taste (scale: 1 to 10)

Ans: Paired Randomization Distribution Design

(c) [1 pt] You have 16 benches in greenhouse, that differ in light and temperature such that each bench has different conditions associated with 4 distinct light and 4 temperature levels: 25%, 50%,75%, and 100% of full sun; 15, 20, 25, 30 C. 4 fertilizer treatments are applied to the benches (75, 150, 225 and 300 mM nitrogen), randomized in such a way that every fertilizer treatment is tested under all 4 light levels, and all 4 temperature levels.

Ans:Latin square design

(d) [1 pt] Suppose that four medical treatments (A,B,C,D) were to be compared at four clinical research sites. But each clinical research site could only collect data on three treatments. The consulting statistician suggested using hospital as a blocking variable.

Ans: Balanced incomplete block design