STAT 5309

R LAB 2

****CONTENTS: 1-FACTOR DESIGN – RESIDUALS CHECKING- TESTS.**

***DUE: Thurs, FEB 7**

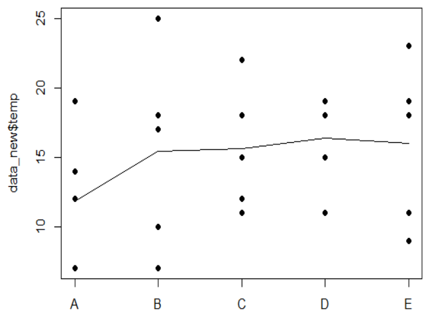
A. PRACTICE

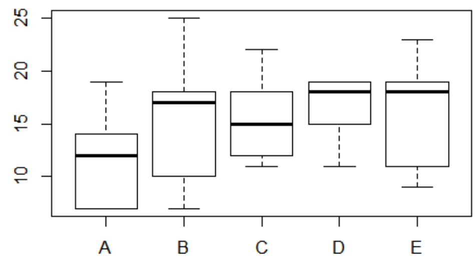
##--------------------Set up a dataframe; observations and levels; ----------

```
# Suppose a factor has 5 levels (called treatment levels or factor levels)
A <- c(7, 7, 15, 11, 9)
B <- c(12, 17, 12, 18, 18)
C <- c(14, 18, 18, 19, 19)
D <- c(19, 25, 22, 19, 23)
E <- c(7, 10, 11, 15, 11)


temp <- c(A,B,C,D,E)              #  combines A,B,C,D,E into a single column vector, length=25
temp

[1]  7  7 15 11  9 12 17 12 18 18 14 18 18 19 19 19 25 22 19 23  7 10 11 15 11

# rep():  repeat a pattern ;  factor():  convert characters into a factor

trt <- rep(c("A", "B", "C", "D", "E"), each=5)  # one column vector, "A"  repeated 5 times, B repeated
5 times..

[1]  A A A A A B B B B B ….. E E E E E


trt <- factor(trt)                                  #make trt into a factor

data_new <- data.frame(trt, temp)                   #combine, as columns, converted into data frame


attach(data.new)
```

#-----------------------------Plots-----------------------------

```
stripchart(data_new$temp ~ data_new$trt, vertical=TRUE,pch=16)

trt_means <- tapply(temp, trt, mean)                #tapply() calculates the treatment means
lines(trt_means)
```





##----------------Linear models: **lm() , anova()---**

#Build a linear model , using lm() or aov()

data.lm <- lm(temp~trt) #a linear model
summary.lm(data.lm) # model summary with C**oefficients**.

```
Call:
lm(formula = temp ~ trt)

Residuals:
   Min     1Q Median     3Q    Max
  -8.4   -4.8    1.6    2.6    9.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.800      2.384   4.949 7.72e-05 ***
trtB           3.600      3.372   1.068    0.298
trtC           3.800      3.372   1.127    0.273
trtD           4.600      3.372   1.364    0.188
trtE           4.200      3.372   1.246    0.227
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 5.331 on 20 degrees of freedom
Multiple R-squared:  0.1076,      Adjusted R-squared:  -0.
07084
F-statistic: 0.6031 on 4 and 20 DF,  p-value: 0.6648
```

NOTES

(a) trtB, trtC, trtD, trtE are **Indicator variables**, created by R, take values { 0,1 }
The control treatment in this case, is Treatment A (R chooses control by alphabetical order). We can choose others as control treatment

(b) Coefficients:
1st coefficient: 11.800 (Intercept) is the mean for Treatment A, which always appears in the model.
2nd coefficient: 3.600 is the difference of Treatment B and Treatment A(can be positive or negative)
To find Treatment B mean, add 11.800+3.600 = 15.400 (as seen in tapply() or stripchart)

```
tapply(temp, trt, mean)                    # tapply()
   A    B    C    D    E
11.8 15.4 15.6 16.4 16.0
```

(c) Change the reference level:

```
trt <- relevel(trt, ref=" B")              # B  is the control level now
```

(b) ANOVA F-test: pay attention to the F-statistic.

➔ **<u>Question 1</u>: Is there significant difference among the factor levels?**

 # --------------summary.lm(); summary.aov();
anova(): compare 2 models (full and reduced model, using F-test)

```
> summary.aov(data.mod)
            Df Sum Sq Mean Sq F value Pr(>F)
trt          4   68.6   17.14   0.603  0.665
Residuals   20  568.4   28.42
```

Notes:

(a) anova() gives same output as **summary.aov()**
(b) Mean Square Error(MSE) : 28.42 (error degree of freedom : 20).
(c)

➔ Question 2:
(a) Check if the SSTotal = SSTreatment + SS Error

(b) Treament df : df(SSTreatment) = a-1 (a is number of factor levels)

(c) Error df: df(Error) = N – a = na – a (n is level repetition)

#--------------------------------Fitted values and Prediction----------------------

```
fitted <-  data.mod$fitted.values
 newdata<- data.frame(trt)
 pred <- predict(data.lm, newdata)
```

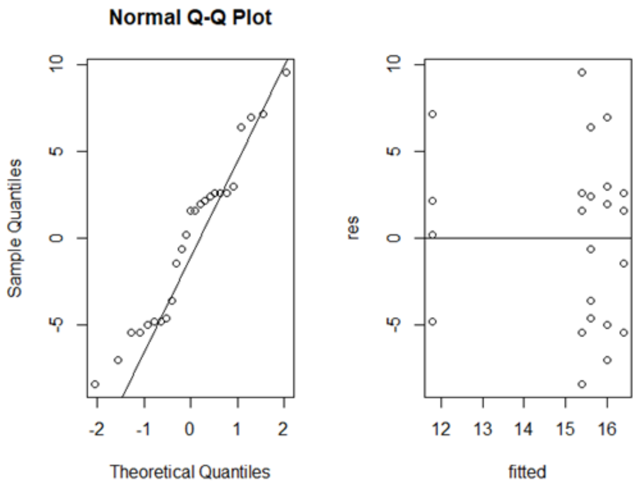
```
> fitted
   1    2    3    4    5    6    7    8    9   10   11   12   13   
14   15   16 
11.8 15.4 15.6 16.4 16.0 11.8 15.4 15.6 16.4 16.0 11.8 15.4 15.6 16
.4 16.0 11.8 
  17   18   19   20   21   22   23   24   25 
15.4 15.6 16.4 16.0 11.8 15.4 15.6 16.4 16.0
```

```
> pred
   1    2    3    4    5    6    7    8    9   10   11   12   13   
14   15   16 
11.8 15.4 15.6 16.4 16.0 11.8 15.4 15.6 16.4 16.0 11.8 15.4 15.6 16
.4 16.0 11.8 
  17   18   19   20   21   22   23   24   25 
15.4 15.6 16.4 16.0 11.8 15.4 15.6 16.4 16.0
```

Note: They are the same (since newdata is just the old data)

##------------------Residuals Checking: Plots, Tests----------------

```
res <- data.mod$residuals

#---------------- qqnorm(), qqline() --------------------
qqnorm(res)
qqline(res)                                   #Check Normality by plots

#---------------- Zero mean/Constant variance of residuals----------------
plot(fitted,res)                              #Check zero mean /constant variance
abline(h=0)                                   #horizontal line through 0
```



#-------------------**shapiro.test() : test Normality of residuals --------------------**

shapiro.test(res)

```
        Shapiro-Wilk normality test

data:  res
W = 0.94816, p-value = 0.2278
```

Note: Null: Normal distributed, Pay attention to P-value

#-------------------- Independence of Residuals : dwtest()------------------

library(lmtest)
dwtest(data.mod, alternative="two.sided") #can use durbin.watson()

```
        Durbin-Watson test

data:  g
DW = 0.77192, p-value = 0.002962
alternative hypothesis: true autocorrelation is not 0
```

Note: NULL: Autocorrelation is 0. Pay attention to P-value.

##-------------------------------------OTHER TESTS--------------------------------

 # -------------Bartlett's test: Test equality of variances among treatment groups:

bartlett.test(temp ~ trt)

```
        Bartlett test of homogeneity of variances

data:  temp by trt
Bartlett's K-squared = 2.0578, df = 4, p-value = 0.725
1
```

---------------**Kruskal-Wallis (test equal treatment means) [Non-parametric]**

kruskal.test(temp ~trt)

```
        Kruskal-Wallis rank sum test

data:  temp and trt
Kruskal-Wallis chi-squared = 2.2842, df = 4, p-value = 0.
6836
```

-------------**Multiple Comparison t-test: Fisher LSD;**

(LSD: Least Significant Difference).

$LSD = t_{\frac{\alpha}{2},N-a}\sqrt{\frac{2MSE}{n}}$. Compare $\left| y_{i.} - y_{j.} \right|$ Against LSD.

MSerror <- 5.331^2
LSD.test(g, "trt", MSerror)
LSD.test(g, "trt", MSerror, console=T)

```
trt,  means and individual ( 95 %) CI

  temp      std r       LCL      UCL Min Max
A 11.8 5.069517 5  6.826825 16.77317   7  19
B 15.4 7.092249 5 10.426825 20.37317   7  25
C 15.6 4.505552 5 10.626825 20.57317  11  22
D 16.4 3.435113 5 11.426825 21.37317  11  19
E 16.0 5.830952 5 11.026825 20.97317   9  23

Alpha: 0.05 ; DF Error: 20
Critical Value of t: 2.085963

least Significant Difference: 7.033131

Treatments with the same letter are not significantly diff
erent.
```

```
  temp groups
D 16.4      a
E 16.0      a
C 15.6      a
B 15.4      a
A 11.8      a
```

#----------- **Multiple comparison test: TukeyHSD()**--------------------------

TukeyHSD(aov(temp~trt), conf.level=0.95)   # the model must be specified

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = temp ~ trt)

$trt
    diff        lwr       upr     p adj
B-A  3.6  -6.489229 13.689229 0.8205602
C-A  3.8  -6.289229 13.889229 0.7905993
D-A  4.6  -5.489229 14.689229 0.6560923
E-A  4.2  -5.889229 14.289229 0.7256590
C-B  0.2  -9.889229 10.289229 0.9999969
D-B  1.0  -9.089229 11.089229 0.9981637
E-B  0.6  -9.489229 10.689229 0.9997545
D-C  0.8  -9.289229 10.889229 0.9992345
E-C  0.4  -9.689229 10.489229 0.9999510
E-D -0.4 -10.489229  9.689229 0.9999510
```

NOTE: Pay attention to the Intervals if they contain 0.

#----------- **Multiple comparison test : Pairwise t-test ()** ---------------

pairwise.t.test(temp, trt)

```
       Pairwise comparisons using t tests with pooled
SD

data:  temp and trt

  A B C D
B 1 - - -
C 1 1 - -
D 1 1 1 -
E 1 1 1 1

P value adjustment method: holm
```

pairwise.t.test(temp,trt, p.adj="bonf")

```
        Pairwise comparisons using t tests with pooled SD

data:  temp and trt

  A B C D
B 1 - - -
C 1 1 - -
D 1 1 1 -
E 1 1 1 1

P value adjustment method: bonferroni
```

B. EXERCISE

1. Data: Bacteria with Packages

Packaging Condition	log(count/cm^2)
Commercial plastic wrap Vacuum packaged 1% CO,40% O2, 59% N 100% CO2	7.66, 6.98, 7.80 5.26, 5.44, 5.80, 7.41, 7.33, 7.04 3.51, 2.91, 3.66

a) Set up the data frame.

(Hint: There are 12 observations . 3 observations for each factor level. Form a vector for factor levels "package". Then form a vector for response, named "logcount". Convert package to factor. Form a data frame, named "bacteria", with "package" and "logcount".

b) Perform a stripchart, with line connecting means, of logcount vs package
c) Build a linear model, using **aov()** response as logcount. Do a **summary.lm()** and summary.aov()
d) Perform a Bartlett test of equal variances.
e) Perform a multiple comparison of treatment mean, using TukeyHSD()

2. Data: Tensile strength of Portland Cement

Four different mixing techniques are used. The following data have be collected.

Mixing Technique	Tensile Strength (lb/in^2)
1	3129 3000 2865 2890

2	3200 3300 2975 3150
3	2800 2900 2985 3050
4	2600 2700 2600 2765

(a) Set up a data frame , with varibles: mixing (factor) and strength (response)
(b) Perform a stripchart. Perform a Box plot.
(c) Use the Fisher LSD (Least Significant Difference) $\alpha = 0.05$ to make comparison

Note: $LSD = t_{\frac{\alpha}{2},N-a}\sqrt{\frac{2MSE}{n}}$

(d) Test the hypothesis that mixing techniques affect the strength of the cement. Use α=0.05 What test do use. Perform the test. Conclusion.

3.

3-6. A manufacturer of television sets is interested in the effect on tube conductivity of four different types of coating for color picture tubes. The following conductivity data are obtained:

Coating Type	Conductivity			
1	143	141	150	146
2	152	149	137	143
3	134	136	132	127
4	129	127	132	129

(a) Is there a difference in conductivity due to coating type? Use $\alpha = 0.05$.
(b) Estimate the overall mean and the treatment effects.
(c) Compute a 95 percent confidence interval estimate of the mean of coating type 4. Compute a 99 percent confidence interval estimate of the mean difference between coating types 1 and 4.
(d) Test all pairs of means using the Fisher LSD method with $\alpha = 0.05$.
(e) Use the graphical method discussed in Section 3-5.3 to compare the means. Which coating type produces the highest conductivity?
(f) Assuming that coating type 4 is currently in use, what are your recommendations to the manufacturer? We wish to minimize conductivity.