stat430 assignment1

Yiming Shen 20891774

27/09/2023

1.10

Firstly, compared with using a sequential approach, a single large experiment is very likely lead to a very large size of data, which requirement strong calculation power to process those data and increase the cost, time, difficulty of experiment.

Secondly, using a sequential approach can reduce the sampling bias. Hence using a single large experiment cannot reduce sampling bias effectively, which could affect the experiment results.

2.27

```
uniformity_125 <- c(2.7,4.6,2.6,3.0,3.2,3.8)
n1 <- length(uniformity_125)
uniformity_200 <- c(4.6,3.4,2.9,3.5,4.1,5.1)
n2 <- length(uniformity_200)
mean1 <- mean(uniformity_125)
mean2 <- mean(uniformity_200)
sd1 <- sd(uniformity_125)
sd2 <- sd(uniformity_200)
df <- n1+n2-2

n1
```

```
## [1] 6
```

```
n2
```

```
## [1] 6
```

```
mean1
```

```
## [1] 3.316667
```

```
mean2
```

```
## [1] 3.933333
```

```
sd1
```

```
## [1] 0.7600439
```

```
sd2
```

```
## [1] 0.821381
```

2.27 (a)

H0: mean1 = mean2 ; HA: mean1 != mean2

```
pooled_sd <- sqrt( ((n1-1)*sd1^2+(n2-1)*sd2^2) / df)
t <- (mean1-mean2) / (pooled_sd*sqrt(1/n1 + 1/n2))

pooled_sd
```

```
## [1] 0.7913069
```

```
t
```

```
## [1] -1.34979
```

```
# two sided test
p_value <- 2*pt(t, df)
p_value
```

```
## [1] 0.2068443
```

Since significance = 0.05 p-value > 0.05 so we do not reject; flow rate does not affect average etch uniformity.

2.27 (b)

Based on 2.27(a), we find that p-value for the test is 0.21

2.27 (c)

H0: sd1^2 / sd2^2 = 1 ; HA: sd1^2 / sd2^2 != 1

```
df1 <- n1-1
df2 <- n2-1
t <- sd1^2 / sd2^2
t
```

```
## [1] 0.8562253
```

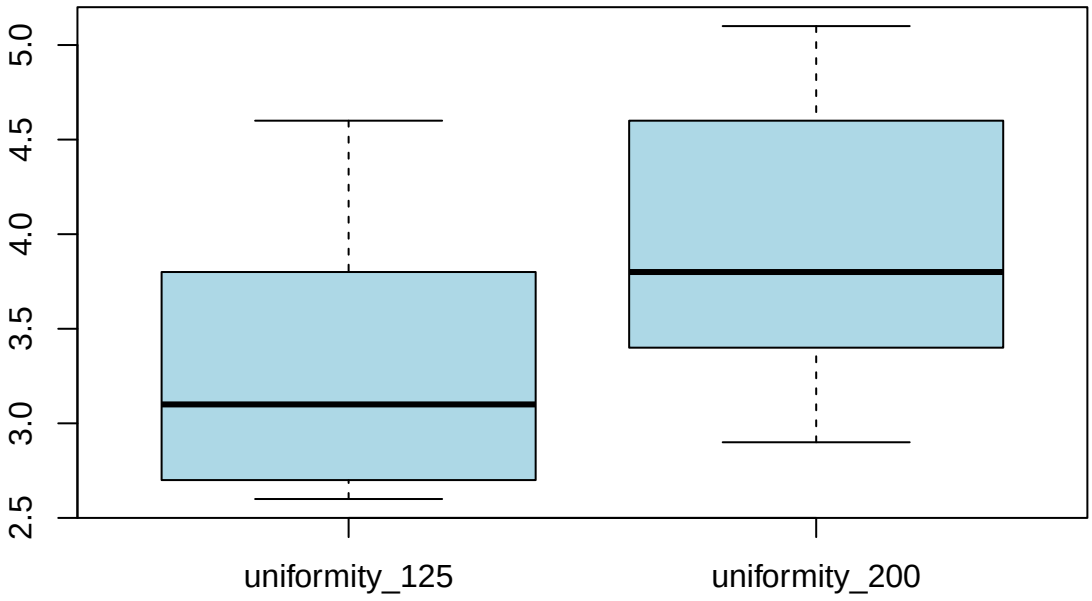
```
p_value <- pf(t,df1,df2) + 1 - pf(1/t,df1,df2)
p_value
```

```
## [1] 0.8689017
```

Since significance level = 0.05 p-value > 0.05 so we do not reject; flow rate does not affect wafer-towafer variability.

2.27 (d)

```
DF <- data.frame(uniformity_125, uniformity_200)
boxplot(DF, col="lightblue")
```



2.29

2.29 (a)

```
thickness_95 <- c(11.176,7.089,8.097,11.739,11.291,10.759,6.467,8.315)
thickness_100 <- c(5.263,6.748,7.461,7.015,8.133,7.418,3.772,8.963)

n1 <- length(thickness_95)
n2 <- length(thickness_100)
mean1 <- mean(thickness_95)
mean2 <- mean(thickness_100)
sd1 <- sd(thickness_95)
sd2 <- sd(thickness_100)
df <- n1 + n2 - 2

n1
```

```
## [1] 8
```

```
n2
```

```
## [1] 8
```

```
mean1
```

```
## [1] 9.366625
```

```
mean2
```

```
## [1] 6.846625
```

```
sd1
```

```
## [1] 2.099564
```

```
sd2
```

```
## [1] 1.640427
```

```
df
```

```
## [1] 14
```

2.29 (a)

H0: mean1 >= mean2 ; HA: mean1 < mean2

```
pooled_sd <- sqrt( ((n1-1)*sd1^2+(n2-1)*sd2^2) / df)
t <- (mean1-mean2) / (pooled_sd*sqrt(1/n1 + 1/n2))

pooled_sd
```

```
## [1] 1.884034
```

```
t
```

```
## [1] 2.675111
```

```
# right-tailed test
p_value <- 1-pt(t, df)
p_value
```

```
## [1] 0.009058979
```

Since significance level = 0.05 p-value < 0.05 we reject; we find higher baking temperature results in lower mean photoresist thickness.

2.29 (b)

Based on 2.29 (a), we find that the p-value for the test is 0.009

2.29 (c)

95% C.I. for difference in mean

```
t <- qt(0.975,df)
# lower bound of 95% C.I.
lb <- (mean1-mean2) - t*pooled_sd*sqrt((1/n1)+(1/n2))
lb
```

```
## [1] 0.4995743
```

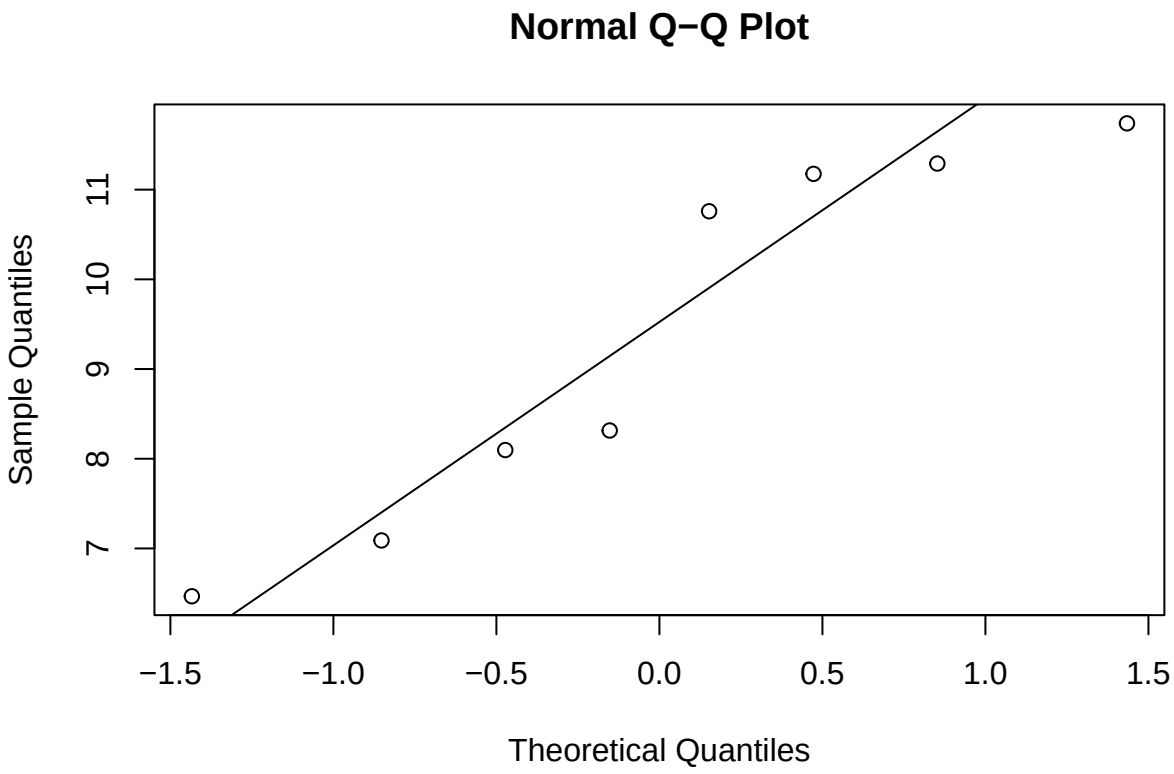
```
# upper bound of 95% C.I.
ub <- (mean1-mean2) + t*pooled_sd*sqrt((1/n1)+(1/n2))
ub
```

```
## [1] 4.540426
```

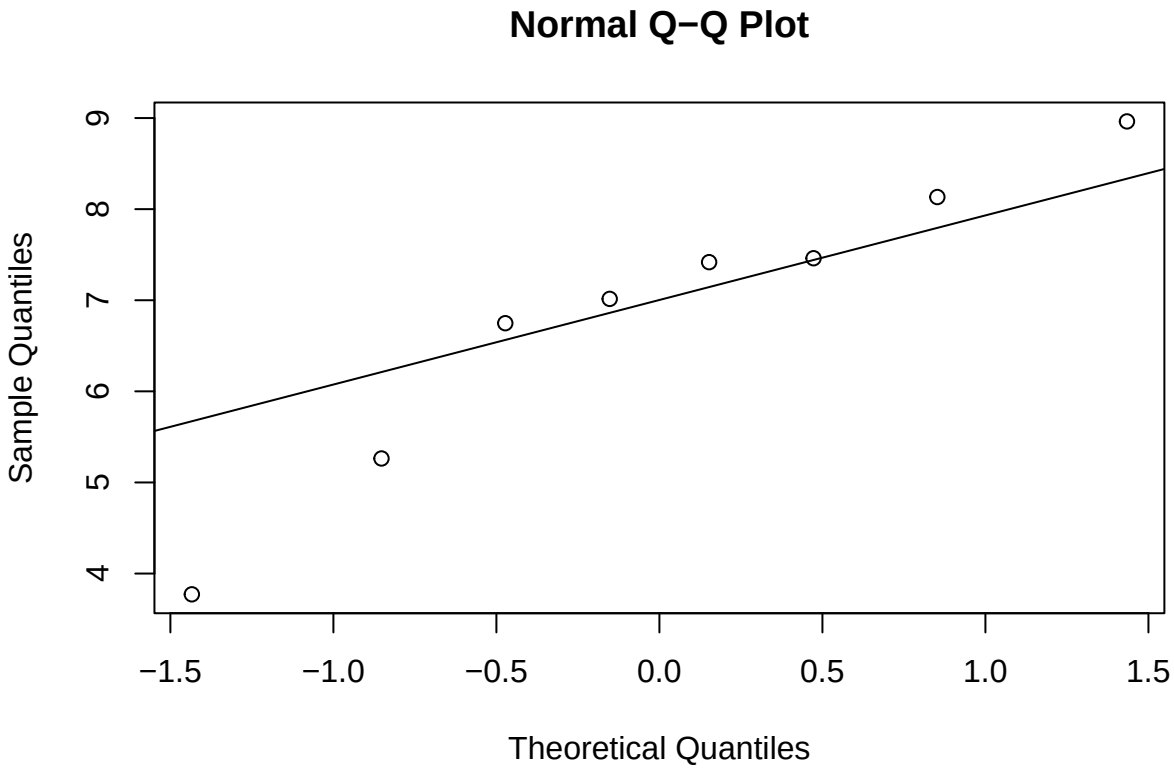
Therefore, the 95% confidence interval for difference of mean is [0.5,4.5] Practical Interpretation: Since zero is not included in this interval, so there exists difference between the photoresist thickness under two different temps.

2.29 (e)

```
# for thickness at 95C
qqnorm(thickness_95)
qqline(thickness_95)
```




```
# for thickness at 100C
qqnorm(thickness_100)
qqline(thickness_100)
```



Based on the QQplots for thickness at 2 different temps, we observe that points spread along the qqline, so we verify the normality assumption.

2.29 (f)

```
power.t.test(n=8,delta=2.5,sd=1,sig.level=.05,power=NULL,type="paired",
             alternative="two.sided")
```

```
##
##      Paired t test power calculation
##
##               n = 8
##           delta = 2.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.9999642
##     alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Therefore, the power is 0.9999

2.29 (g)

```
power.t.test(n=NULL,delta=1.5,sd=1,sig.level=.05,power=0.9,type="paired",
             alternative="two.sided")
```

```
##
##      Paired t test power calculation
##
##               n = 6.869959
##           delta = 1.5
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Therefore, sample size is n=roof(6.8699)=7 pairs

3.15

```
contribution <- c(1000,1500,1200,1800,1600,1100,1000,1250,
                  1500,1800,2000,1200,2000,1700,1800,1900,
                  900,1000,1200,1500,1200,1550,1000,1100)

approach <- c(rep(1,8), rep(2,8), rep(3,8))
DF <- data.frame(contribution, approach)
m <- 3
n <- length(DF$contribution)
approach1 <- DF$contribution[DF$approach==1]
approach2 <- DF$contribution[DF$approach==2]
approach3 <- DF$contribution[DF$approach==3]
approach1
```

```
## [1] 1000 1500 1200 1800 1600 1100 1000 1250
```

```
approach2
```

```
## [1] 1500 1800 2000 1200 2000 1700 1800 1900
```

```
approach3
```

```
## [1]  900 1000 1200 1500 1200 1550 1000 1100
```

```
mean1 <- mean(approach1)
mean2 <- mean(approach2)
mean3 <- mean(approach3)
grand_mean <- mean(DF$contribution)
mean1
```

```
## [1] 1306.25
```

```
mean2
```

```
## [1] 1737.5
```

```
mean3
```

```
## [1] 1181.25
```

```
grand_mean
```

```
## [1] 1408.333
```

3.15 (a)

H0: mean1 = mean2 = mean3 ; HA: there exists at least one inequality

```
# using ANOVA test

n1 <- n/m  # = 24 sponsors / 3 approaches
n2 <- n1
n3 <- n2

SSc <- n1*(mean1-grand_mean)^2 + n2*(mean2-grand_mean)^2 + n3*(mean3-grand_mean)^2
SSc
```

```
## [1] 1362708
```

```
DFc <- m-1
MSc <- SSc / DFc
diff1_sqr <- (approach1 - mean1)^2
diff2_sqr <- (approach2 - mean2)^2
diff3_sqr <- (approach3 - mean3)^2
SSe <- sum(diff1_sqr) + sum(diff2_sqr) + sum(diff3_sqr)
DFe <- n-m
MSe <- SSe / DFe

t <- MSc / MSe
t
```

```
## [1] 9.409577
```

```
# use one-tailed test
p_value <- 1-pf(t,DFc,DFe)
p_value
```

```
## [1] 0.00120875
```

Since signifiance level is 0.05 p-value < 0.05 we reject; Data indicates that there is a difference in results with 3 different approaches.

3.15 (b)

```
residual1 <- approach1 - mean1
residual1
```

```
## [1] -306.25  193.75 -106.25  493.75  293.75 -206.25 -306.25  -56.25
```

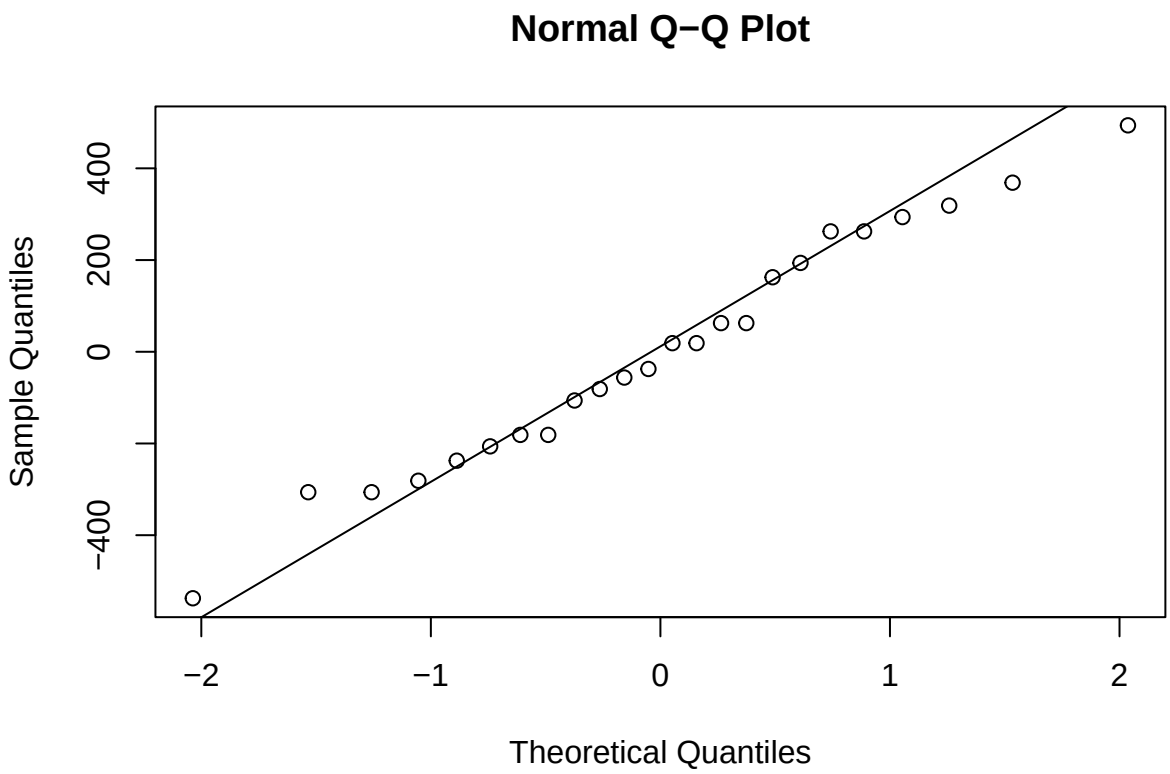
```
residual2 <- approach2 - mean2
residual2
```

```
## [1] -237.5   62.5  262.5 -537.5  262.5  -37.5   62.5  162.5
```

```
residual3 <- approach3 - mean3
residual3
```

```
## [1] -281.25 -181.25   18.75  318.75   18.75  368.75 -181.25  -81.25
```

```
residual <- c(residual1, residual2, residual3)
# Normality check
qqnorm(residual)
qqline(residual)
```



Based on the QQplot, we verify the normality of residuals, which shows the model adequacy.

3.15 (c)

```
# Bartlett's test
k <- 3
n
```

```
## [1] 24
```

```
sd1 <- sd(approach1)
sd2 <- sd(approach2)
sd3 <- sd(approach3)

pooled_sd_sqr <- ( (n1-1)*sd1^2 + (n2-1)*sd2^2 + (n3-1)*sd3^2 ) / (n-k)
pooled_sd_sqr
```

```
## [1] 72410.71
```

```
# Numerator of Test Stat
t_num <- (n-k)*log(pooled_sd_sqr) -
  ( (n1-1)*log(sd1^2) + (n2-1)*log(sd2^2) + (n3-1)*log(sd3^2) )

# Denominator of Test Stat
t_den <- 1 + (1/(3*(k-1)))*( (1/(n1-1))+(1/(n2-1))+(1/(n3-1))-(1/(n-k)) )

t <- t_num / t_den
t
```

```
## [1] 0.335604
```

```r
p_value <- 1-pchisq(t,k-1)
p_value
```

```
## [1] 0.8455212
```

Since significance is 0.05 p-value > 0.05 We do not reject the null of equal variance.

3.15 (d)

```r
# Tukey's LSD

# Step 1
# ANOVA has been conducted in (a)

# Step 2
# critical number
df <- n-k
df
```

```
## [1] 21
```

```r
c <- qt(.975,df)
c
```

```
## [1] 2.079614
```

```r
# Step 3
# t-scores
SE <- sqrt(MSe / n1)
score1 <- (mean2 - mean1)/SE
score1
```

```
## [1] 4.532864
```

```r
score2 <- (mean1 - mean3)/SE
score2
```

```
## [1] 1.313874
```

```r
score3 <- (mean2 - mean3)/SE
score3
```

```
## [1] 5.846738
```

Therefore, we find: score1 larger than c ; score2 smaller than c ; score3 larger than c .

we can say: approach 1 is significantly different than 2 ; approach 2 is significantly different than 3 ; approach 1 is not sig. diff. than approach 3

3.15 (e)

Ho: 2mean1 = mean2 + mean3 / 2mean1 - (mean2 + mean3) = 0 ; HA: equation unvaild

```r
# contrasts - 1 group vs. 2 groups
contrast <- 2*mean1 - (mean2 + mean3)
c <- c(2,-1,-1)
var.contr <- MSe/n * sum(c^2)

t <- contrast / var.contr
t
```

```
## [1] -0.01691739
```

```
p_value <- 1-pt(t,df)
p_value
```

```
## [1] 0.5066689
```

Since significance level is 0.05 p-value > 0.05 we do not reject the null hypothesis that 2mean1 = mean2 + mean3

3.18

```
conductivity <- c(143,141,150,146,
                  152,149,137,143,
                  134,136,132,127,
                  129,127,132,129)

type <- c(rep(1,4), rep(2,4), rep(3,4), rep(4,4))

df <- data.frame(type, conductivity)
type1 <- df$conductivity[df$type==1]
type2 <- df$conductivity[df$type==2]
type3 <- df$conductivity[df$type==3]
type4 <- df$conductivity[df$type==4]
n <- length(df$conductivity)
n1 <- length(type1)
n2 <- n1
n3 <- n2
n4 <- n3

mean1 <- mean(type1)
mean2 <- mean(type2)
mean3 <- mean(type3)
mean4 <- mean(type4)
grand_mean <- mean(df$conductivity)
```

3.18(a)

H0: mean1 = mean2 = mean3 = mean4 ; HA: at least one inequality

```
m <- 4
SSc <- n1*(mean1-grand_mean)^2 + n2*(mean2-grand_mean)^2 +
       n3*(mean3-grand_mean)^2 + n4*(mean4-grand_mean)^2
SSc
```

```
## [1] 844.6875
```

```
DFc <- m-1
MSc <- SSc / DFc
diff1_sqr <- (type1 - mean1)^2
diff2_sqr <- (type2 - mean2)^2
diff3_sqr <- (type3 - mean3)^2
diff4_sqr <- (type4 - mean4)^2
SSe <- sum(diff1_sqr) + sum(diff2_sqr) + sum(diff3_sqr) + sum(diff4_sqr)
DFe <- n-m
MSe <- SSe / DFe

t <- MSc / MSe
t
```

```
## [1] 14.30159
```

```
# use one-tailed test
p_value <- 1-pf(t,DFc,DFe)
p_value
```

```
## [1] 0.0002881237
```

Since significance level is 0.05 p-value < 0.05 we reject; there is difference in conductivity due to coating type

3.18 (b)

```
grand_mean
```

```
## [1] 137.9375
```

- estimate the overall mean: grand_mean = 137.9375

```
mean1 - grand_mean
```

```
## [1] 7.0625
```

- estimate type 1 's treatment effect: mean1 - grand_mean = 7.0625

```
mean2 - grand_mean
```

```
## [1] 7.3125
```

- estimate type 2 's treatment effect: mean2 - grand_mean = 7.3125

```
mean3 - grand_mean
```

```
## [1] -5.6875
```

- estimate type 3 's treatment effect: mean3 - grand_mean = -5.6875

```
mean4 - grand_mean
```

```
## [1] -8.6875
```

- estimate type 4 's treatment effect: mean4 - grand_mean = -8.6875
- Using type as a 'left out' category

3.18 (c)

```
sd1 <- sd(type1)
sd1
```

```
## [1] 3.91578
```

```
sd2 <- sd(type2)
sd2
```

```
## [1] 6.652067
```

```
sd3 <- sd(type3)
sd3
```

```
## [1] 3.86221
```

```
sd4 <- sd(type4)
sd4
```

```
## [1] 2.061553
```

```
df <- n-m
pooled_sd <- sqrt( ((n1-1)*sd1^2+(n2-1)*sd2^2
                    +(n3-1)*sd3^2+(n4-1)*sd4^2) / df)
pooled_sd_sqr <- pooled_sd^2
pooled_sd_sqr
```

```
## [1] 19.6875
```

```
# 95% C.I. for the mean of type 4:
# we got constant c = 2.179
c <- 2.179
# lower bound
lb <- mean4 - c*sqrt(pooled_sd_sqr/n4)
# upper bound
ub <- mean4 + c*sqrt(pooled_sd_sqr/n4)
lb
```

```
## [1] 124.4158
```

```
ub
```

```
## [1] 134.0842
```

```
# 99% C.I. for the mean difference between type 1 and 4
# we find constant  c = 3.055
c <- 3.055
# lower bound
lb <- (mean1-mean4)-c*sqrt((2*pooled_sd_sqr)/((n1+n2)/2))
ub <- (mean1-mean4)+c*sqrt((2*pooled_sd_sqr)/((n1+n2)/2))
lb
```

```
## [1] 6.165014
```

```
ub
```

```
## [1] 25.33499
```

Therefore, 95% for type4 mean is [124.4158,134.0842] 99% for mean difference between type 1 and 4 is [6.165014,25.33499]

3.18 (d)

```
# Tukey's LSD

# Step 1
# ANOVA has been conducted in (a)

# Step 2
# critical number
c <- qt(.975,df)
c
```

```
## [1] 2.178813
```

```
# Step 3
# t-scores
SE <- sqrt(MSe / n1)
score1 <- (mean2 - mean1)/SE
score1
```

```
## [1] 0.1126872
```

```
score2 <- (mean1 - mean3)/SE
score2
```

```
## [1] 5.747049
```

```
score3 <- (mean1 - mean4)/SE
score3
```

```
## [1] 7.099296
```

```
score4 <- (mean2 - mean3)/SE
score4
```

```
## [1] 5.859736
```

```
score5 <- (mean2 - mean4)/SE
score5
```

```
## [1] 7.211983
```

```
score6 <- (mean3 - mean4)/SE
score6
```

```
## [1] 1.352247
```

Therefore, we find: score1 smaller than c ; score2 larger than c ; score3 larger than c ; score4 larger than c ; score5 larger than c ; score6 smaller than c ;

we can say: approach 1 is significantly different than 3 ; approach 1 is significantly different than 4 ; approach 2 is significantly different than 3 ; approach 2 is significantly different than 4 ; approach 1 is not sig. diff. than approach 3 ;
approach 1 is not sig. diff. than approach 2

3.18 (f)

Based on data, type 3 and 4 both have low mean values of conductivity. Since type 4 is currently in use, so I would recommend that not change the coating type or change to type 3.