STAT500: Home_Work_2

Zebosi Brian, Zhengqiang Ni, Ruina Chang

9/20/2021

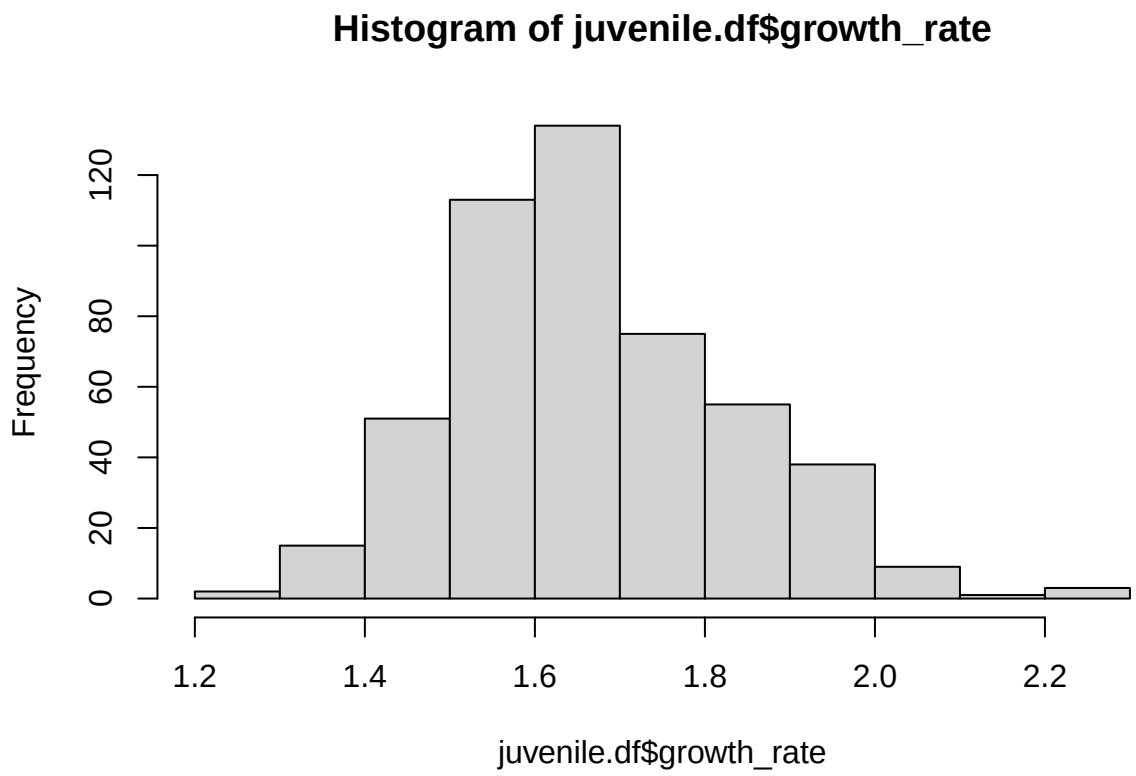
Question_1

a). Plot a histogram of the variable growth_rate

```
library(fishdata)
```

```
## Warning: package 'fishdata' was built under R version 4.0.5
```

```
juvenile.df <- juvenile_metrics
hist(juvenile.df$growth_rate)
```



Growth data doesn't seem to be normally distributed. i.e. data tends be skewed to slightly skewed.

b). Use the function ks.test in R to test a normality

```
ks.test(juvenile.df$growth_rate, 'pnorm',
        mean(juvenile.df$growth_rate),
```

```
        sd(juvenile.df$growth_rate))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  juvenile.df$growth_rate
## D = 0.074733, p-value = 0.00785
## alternative hypothesis: two-sided
```

p-value = 0.00785, reject the null hypothesis that the growth rate data follows a normal distribution.

i.e. Accept the alternative hypothesis that growth rate data doesnot follow a normal distribution.

c).

```
growth.log <- log(juvenile.df$growth_rate)

t.test(growth.log, mu = 1.7)
```

```
##
##  One Sample t-test
##
## data:  growth.log
## t = -275.64, df = 495, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 1.7
## 95 percent confidence interval:
##  0.4977464 0.5147647
## sample estimates:
## mean of x
## 0.5062556
```

```
exp(1.7-(mean(growth.log)))
```

```
## [1] 3.299413
```

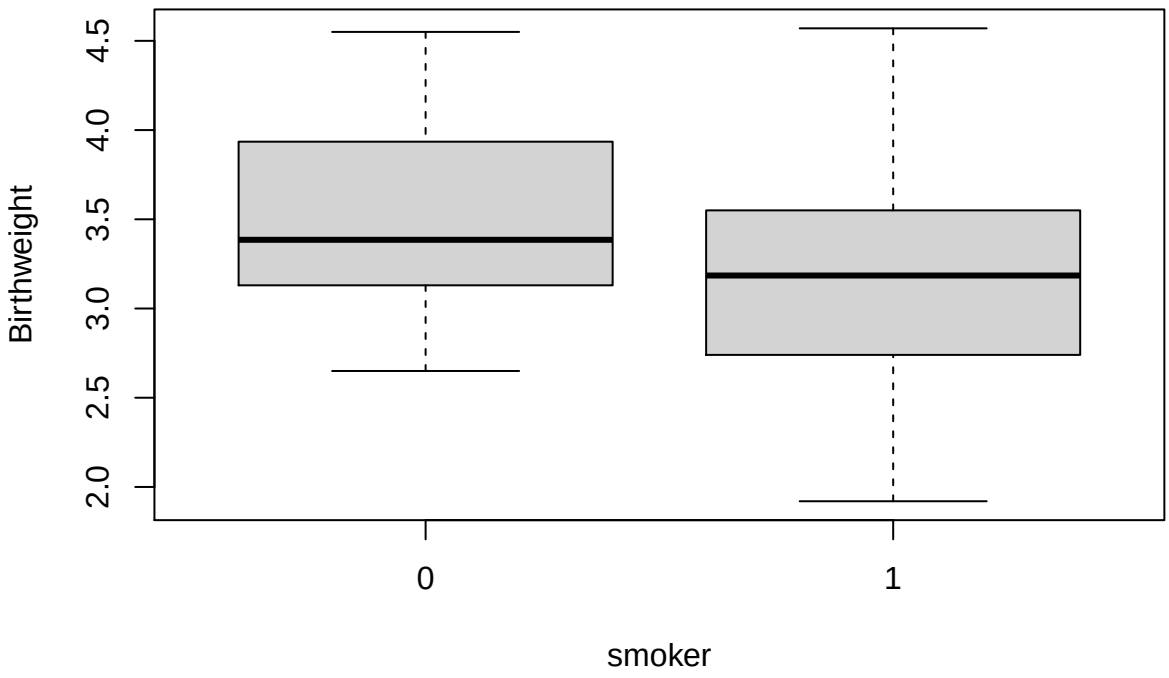
growth rate is 3.2 times less than median growth rate.

Question_2

a)

```
bw.df <- read.csv("Birthweight_reduced_kg_R.csv")
```

```
boxplot(Birthweight~smoker, data=bw.df)
```



T-test using calculated manaully

```
m1 <- length(bw.df$Birthweight[bw.df$smoker == 1])
m2 <- length(bw.df$Birthweight[bw.df$smoker == 0])
s2 <- var(bw.df$Birthweight[bw.df$smoker == 0])
s1 <- var(bw.df$Birthweight[bw.df$smoker == 1])
Sp2 <- (s1+s2)/2
test.stat <- (mean(bw.df$Birthweight[bw.df$smoker == 1])-
                (mean(bw.df$Birthweight[bw.df$smoker == 0])))/sqrt(Sp2*(1/m1+1/m2))

test.stat
```

```
## [1] -2.103574
```

```
y <- ((m1+m2)-2)
2*(1-pt(abs(test.stat), y))
```

```
## [1] 0.04175107
```

T-test using R t.test function

```
t.test(bw.df$Birthweight[bw.df$smoker == 0], bw.df$Birthweight[bw.df$smoker == 1],
       conf.level = 0.90, alternative = "two.sided",var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  bw.df$Birthweight[bw.df$smoker == 0] and bw.df$Birthweight[bw.df$smoker == 1]
## t = 2.0934, df = 40, p-value = 0.0427
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.0734489 0.6773693
## sample estimates:
## mean of x mean of y
##  3.509500  3.134091
```

c).

```
mean_diff = mean(bw.df$Birthweight[bw.df$smoker == 0])- (mean(bw.df$Birthweight[bw.df$smoker == 1]))
s.e.m <- sqrt(Sp2*(1/m1+1/m2))
qtn <- abs(qt(0.05,((m1+m2)-2)))
Lower_cl <- mean_diff-(qtn*s.e.m)
Lower_cl
```

```
## [1] 0.07490483
```

```
Upper_cl <- mean_diff+(qtn*s.e.m)
Upper_cl
```

```
## [1] 0.6759134
```

manual computation:(0.074 , 0.675) and using R-t.test function (0.0734 , 0.677)

d).

P < 0.05, suggests that there is moderate evidence for difference between mother's smoking habits.

We are 90% confident that the difference between mother's smoking habits is 0.0734 between 0.677.

Question_3

```
chol.df <- read.csv("Cholesterol_R.csv")
```

a)

Experimental unit : 18 individuals diagnosed with high cholesterol who replaced butter

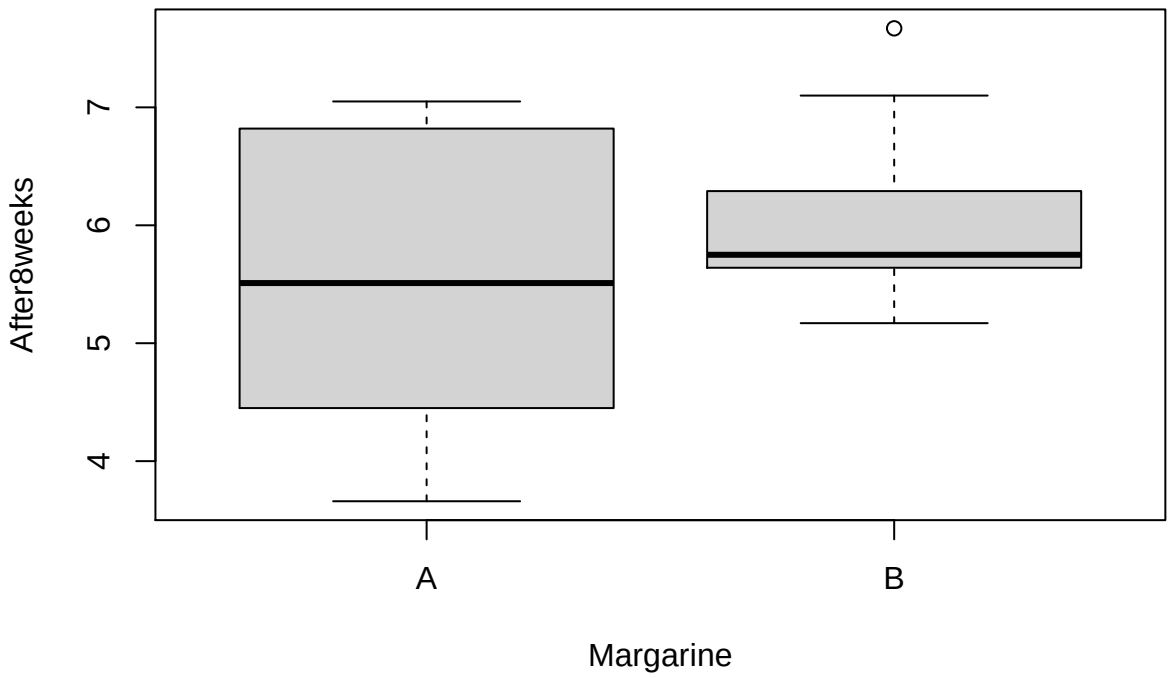
Population : individuals diagnosed with high cholesterol who replaced butter

Treatments : margarine [A and B]

Response : blood cholesterol levels

b)

```
boxplot(After8weeks~Margarine, data = chol.df)
```



C)

Computation of t.test by hand

```
m1 <- length(chol.df$After8weeks[chol.df$Margarine == "A"])
m2 <- length(chol.df$After8weeks[chol.df$Margarine == "B"])
s2 <- var(chol.df$After8weeks[chol.df$Margarine == "A"])
s1 <- var(chol.df$After8weeks[chol.df$Margarine == "B"])
Sp2 <- (s1+s2)/2
test.stat <- (mean(chol.df$After8weeks[chol.df$Margarine == "A"])-
                (mean(chol.df$After8weeks[chol.df$Margarine == "B"])))/sqrt(Sp2*(1/m1+1/m2))

test.stat
```

```
## [1] -1.125284
```

```
df <- ((m1+m2)-2)
2*(1-pt(abs(test.stat), df))
```

```
## [1] 0.2770664
```

T.test using R_t.test function

```
t.test(chol.df$After8weeks[chol.df$Margarine == "B"], chol.df$After8weeks[chol.df$Margarine == "A"],
       conf.level = 0.95, alternative = "two.sided",var.equal = T)
```

```
##
##  Two Sample t-test
##
```

```
## data:  chol.df$After8weeks[chol.df$Margarine == "B"] and chol.df$After8weeks[chol.df$Margarine == "A
## t = 1.1253, df = 16, p-value = 0.2771
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5126535  1.6726535
## sample estimates:
## mean of x mean of y
##  6.068889  5.488889
```

d)

(-0.5126535, 1.6726535)

e)

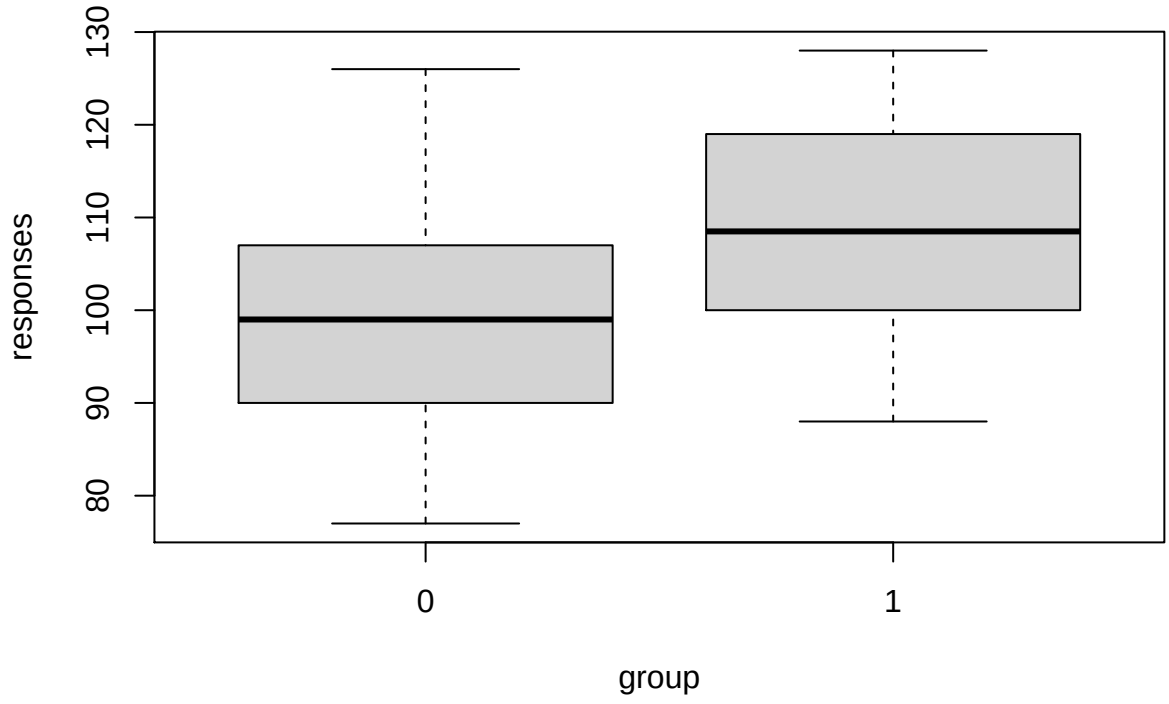
P-value > 0.05, there is little or no evidence that there is difference in mean cholesterol reduction between the two brands of margarine after 8 weeks of use

Question_4

```
att.df <- read.csv("attendance.csv")
```

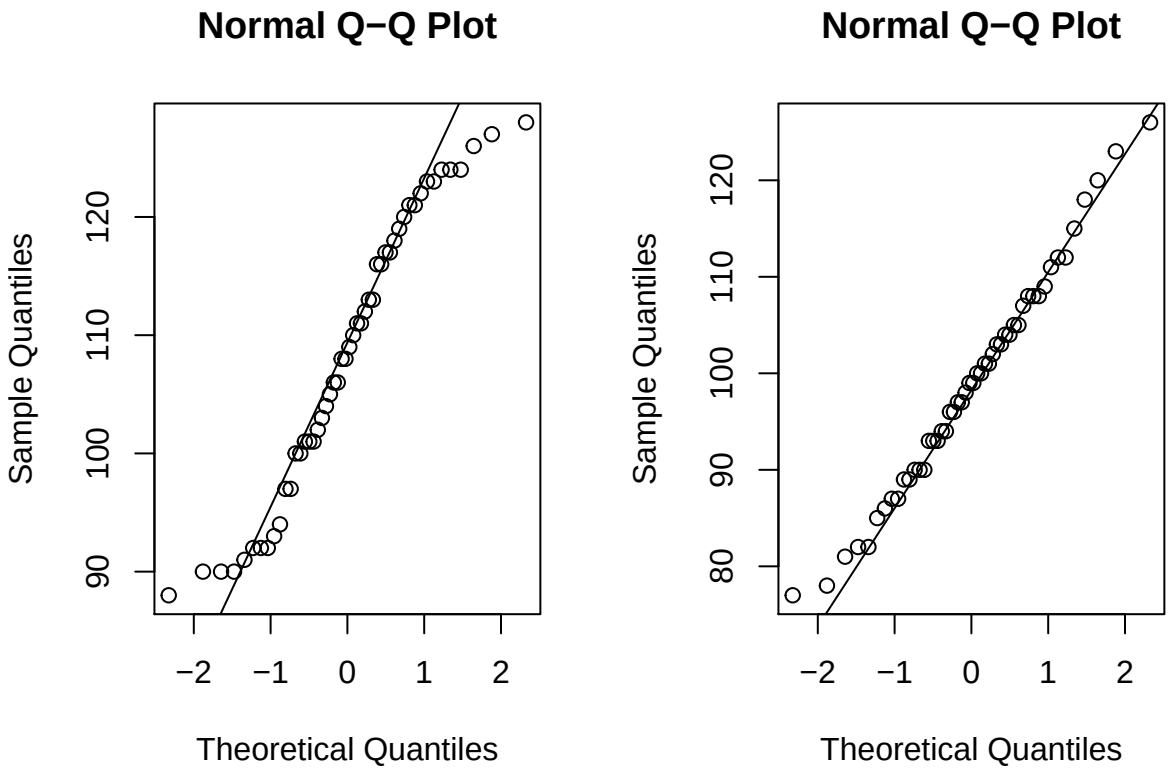
a)

```
boxplot(responses~group, data=att.df)
```



b)

```
par(mfrow=c(1,2))
qqnorm(att.df$responses[att.df$group == 1])
qqline(att.df$responses[att.df$group == 1])
qqnorm(att.df$responses[att.df$group == 0])
qqline(att.df$responses[att.df$group == 0])
```

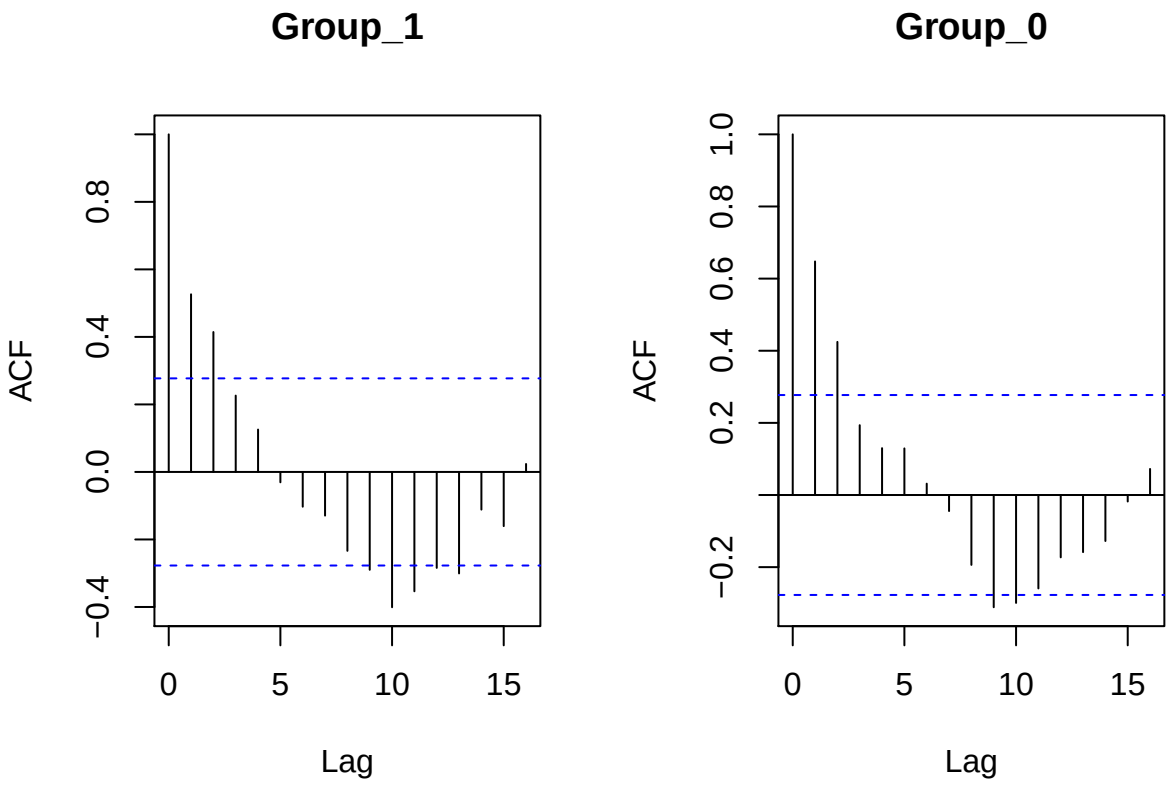


Plots of the data look close to random samples from a normal distribution.

seem normally distributed.

c).

```
par(mfrow=c(1,2))
acf(att.df$responses[att.df$group == 1], main="Group_1")
acf(att.df$responses[att.df$group == 0], main="Group_0")
```



d)

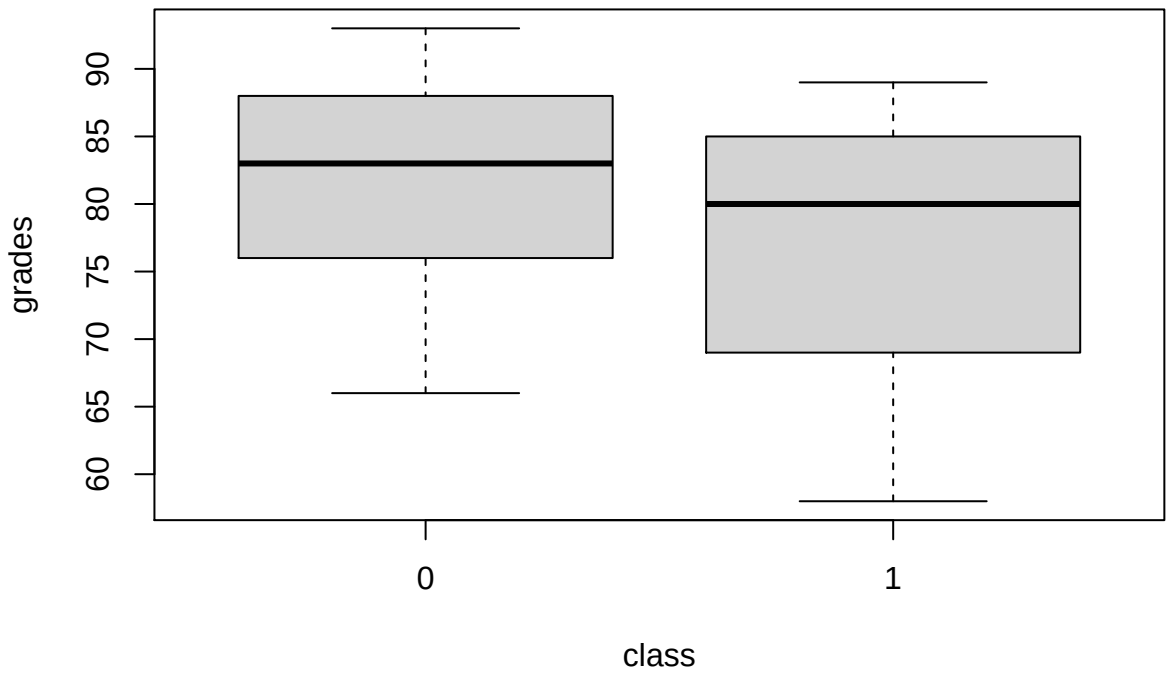
T.test wouldn't adequately account for the variation because there is correlation between points i.e. Assumption of independence is violated.

Question_5

```
grad.df <- read.csv("grades.csv")
```

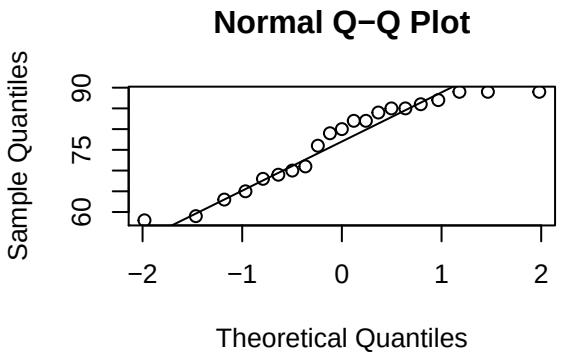
a)

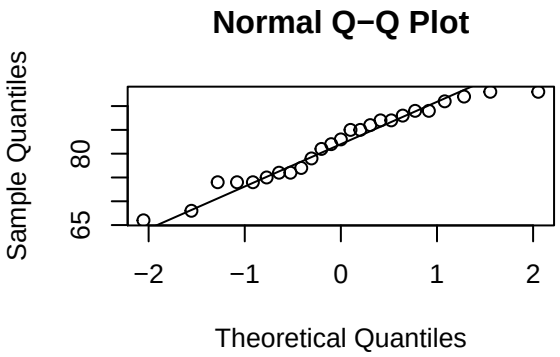
```
boxplot(grades~class, data=grad.df)
```

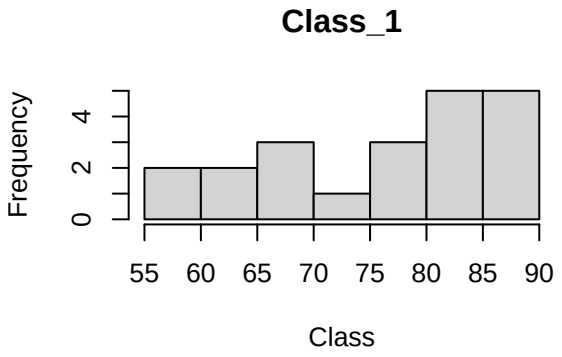


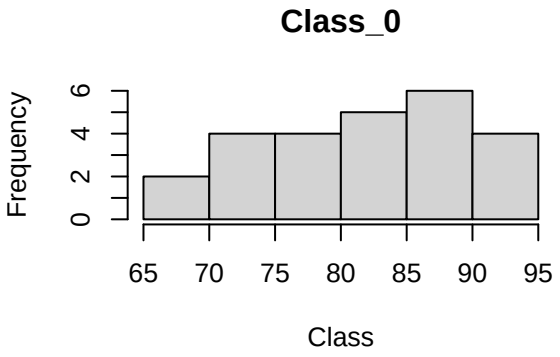
b)

```
par(mfrow=c(2,2))
qqnorm(grad.df$grades[grad.df$class== 1])
qqline(grad.df$grades[grad.df$class== 1])
qqnorm(grad.df$grades[grad.df$class== 0])
qqline(grad.df$grades[grad.df$class== 0])
hist(grad.df$grades[grad.df$class== 1], xlab="Class",main="Class_1")
hist(grad.df$grades[grad.df$class== 0], xlab="Class",main="Class_0")
```



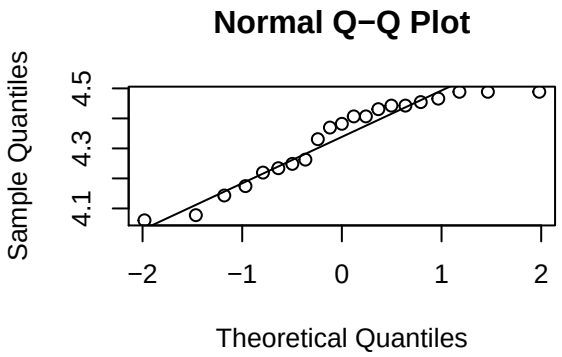


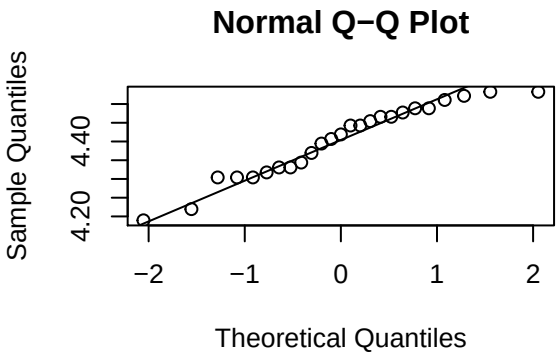


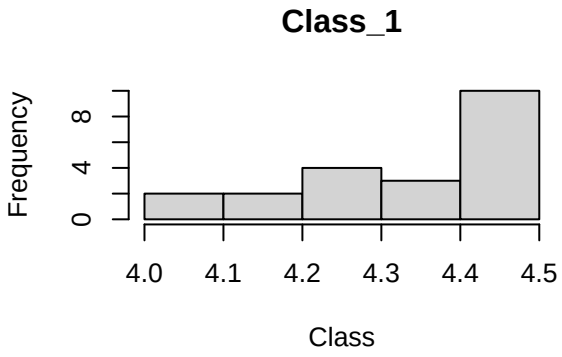


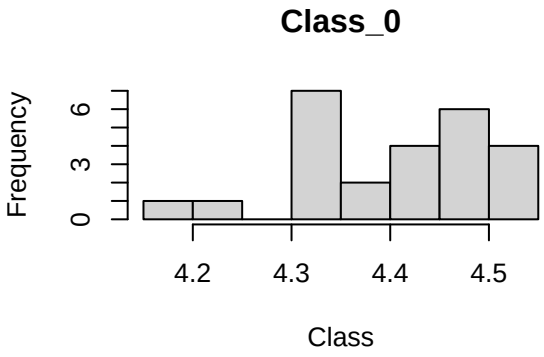
c)

```
grad.df$log_grades <- log(grad.df$grades)
par(mfrow=c(2,2))
qqnorm(grad.df$log_grades[grad.df$class== 1])
qqline(grad.df$log_grades[grad.df$class== 1])
qqnorm(grad.df$log_grades[grad.df$class== 0])
qqline(grad.df$log_grades[grad.df$class== 0])
hist(grad.df$log_grades[grad.df$class== 1], xlab="Class",main="Class_1")
hist(grad.df$log_grades[grad.df$class== 0], xlab="Class",main="Class_0")
```









c)

log transformation of grades didn't improve the normality of the grades.

d)

```
wilcox.test(grad.df$grades[grad.df$class == 1], grad.df$grades[grad.df$class == 0],
            conf.level = 0.9, alternative = "two.sided")
```

```
## Warning in wilcox.test.default(grad.df$grades[grad.df$class == 1],
## grad.df$grades[grad.df$class == : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  grad.df$grades[grad.df$class == 1] and grad.df$grades[grad.df$class == 0]
## W = 190.5, p-value = 0.1143
## alternative hypothesis: true location shift is not equal to 0
```