Homework 6

Teryl Schmidt | tschmidt6@wisc.edu | 9072604920 | Discussion 312 | Grader: Chi-Shain Dai
11/1/2018

Problem 1

An eel supply company has two large vats of adolescent eels (called 'elvers'). One vat contains American eels, and the other contains European eels. An eel researcher suspects that American eels have a larger number of scales than European eels. Let μAm be the population mean for American, and μEu be the population mean for European. The species of eels can be considered independent. The researcher visits the company and uses an ingenious method to get random samples from each vat (details omitted) and counts the number of scales on each selected elver. Suppose the data from each sample supports the assumption that both species scale counts are approximately normal. The sample statistics are summarized below:
Type | Sample Size | Sample Mean | Sample Variance |
American | 8 | 220 | 21 |
European | 5 | 204 | 23 |

a. Explain why a two sample t test with an equal variance assumption is appropriate to use here. List the null and alternative hypothesis.

The two samples are independent, and normally distributed. The variances are close. So all the three assumptions are met for 2-sample t-test.
Ho: $\mu_{US} - \mu_{EU} = 0$
Ha: $\mu_{US} - \mu_{EU} > 0$

b. At α = 0.1, find the rejection region in the scale of t and also in the scale of X¯Am − X¯Eu and use them to make a reject or not reject decision based on the observed test statistic and observed difference in means, respectively. Then make a conclusion in the context of the problem.

The rejection region of the scaleof t is: $t_{((8+5)-2\ ,\ 0.1)} = 1.36$
Based on t = X¯Am − X¯Eu / s_p * sqrt(1 / n_{Am} + 1 / n_{Eu})
to reverse standardize we need to find the estimated standard error.
SE = sqrt(((7 * 21) + (4 * 23)) / 11) = 4.661252
Reverse standardize get 1.36 * 4.661252 = 6.339303
The rejection region is reject if X¯Am − X¯Eu > 6.339303
220 - 204 = 16
16 > 6.339303 so reject the null hypothesis.

c. Decide the same test based on the p-value compared to α = 0.1.

SE = sqrt(((7 * 21) + (4 * 23)) / 11) = 4.661252 t_{obs} = (220 - 204) / (4.66 * sqrt(1/8 + 1/5)) = 6.022716
pvalue = P(t_{12} > 6.022716)
pvalue < 0.0001
The null hypothesis is rejected. We conclude that the population mean for American is different from the population mean for European.

d. Suppose we instead wanted to test: H0 : μAm − μEu = 10 vs. HA : μAm − μEu > 10 If the same data is used for both tests, would the p-value for this test be larger or smaller than the p-value that would have been computed for the test in part (a)? Explain your answer. You do not need to actually compute either p-value to answer this question.

The pvalue would be bigger because 16 is closer to 10 than 0 is.

e. Suppose instead we choose to perform a Welch's t test. What does that mean for the assumptions we are making? Perform a Welch's T test for the same data. How much do the p values differ?

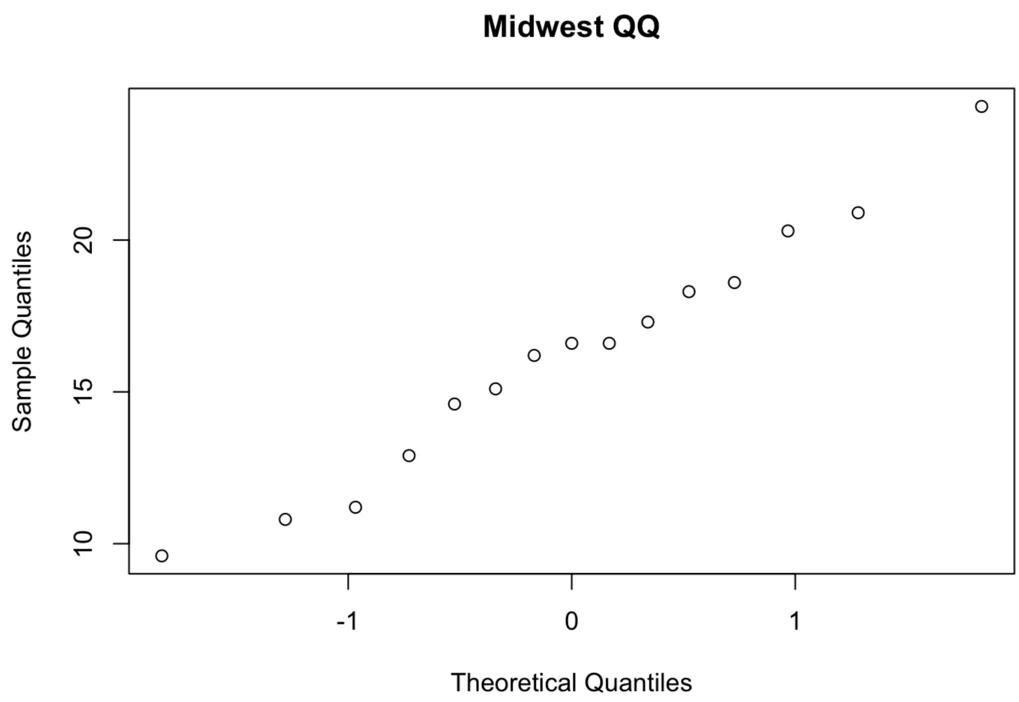
t_v = X¯Am − X¯Eu - 0 / sqrt(S^2_{Am} / n_{Am} + S^2_{Eu} / n_{Eu})
= (220 - 204 - 0) / sqrt((21 / 8) + (23 / 5)) = 5.952523
v = (220 - 204 - 0) / ((21 / 8) + (23 / 5))^2 / ((21 / 8)^2 + (23 / 5)$^{2)}$2 = 241.1818 P(t_{241} > 5.952523) < 0.0005
= pvalue < 0.0005
We assume population and samples are independent, and that the data is normal. We do not assume equal variance. Since pvalue < 0.0005 there is strong evidence to suggest that the population mean for American is different from the population mean for European.

Problem 2

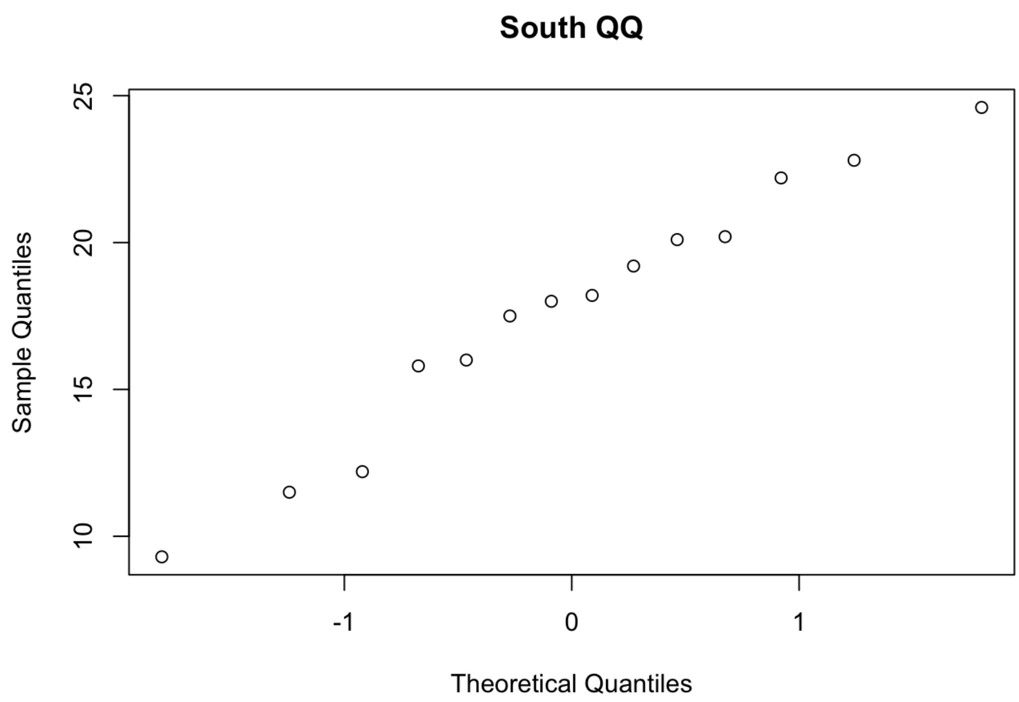
Data on household vehicle miles of travel (VMT) are compiled annually by the Federal Highway Administration. A researcher is interested in whether there is a difference in last year's mean MVT for midwestern and southern households. Independent random samples of 15 midwestern households and 14 southern households provided the following data on last year's VMT, in thousands of miles:
Midwest : 16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3
South : 22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5

a. Graph the data as you see fit. Why did you choose the graph(s) you did and what does it (do they) tell you? Also calculate summary statistics relevant to the research question.

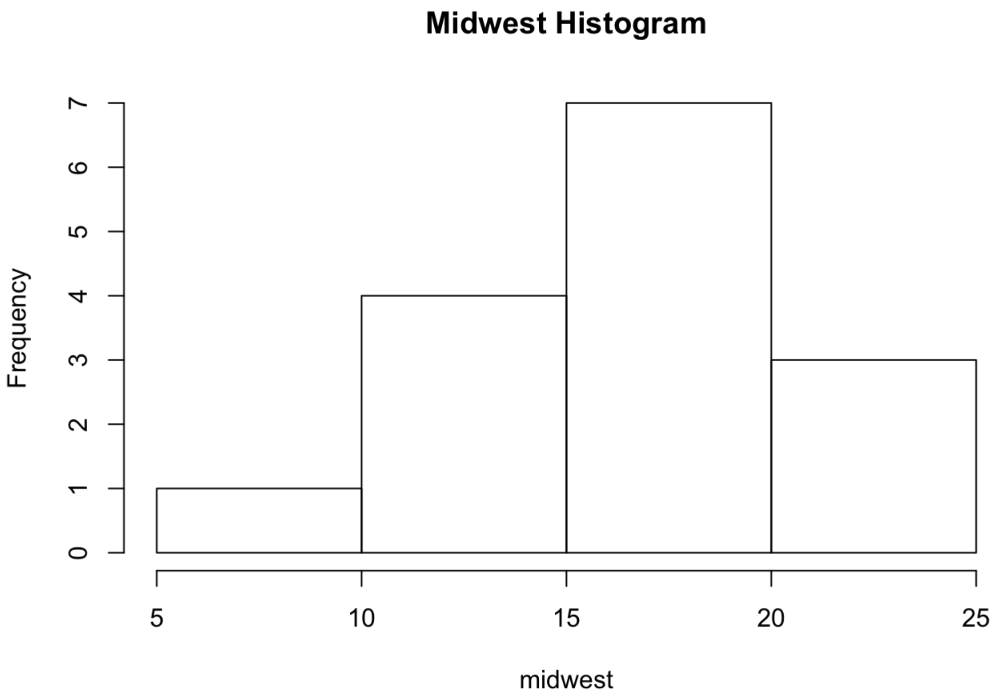
```
midwest = c(16.2, 12.9, 17.3, 14.6, 18.6, 10.8, 11.2, 16.6, 16.6, 24.4, 20.3, 20.9, 9.6, 15.1, 18.3)
south = c(22.2, 19.2, 9.3, 24.6, 20.2, 15.8, 18.0, 12.2, 20.1, 16.0, 17.5, 18.2, 22.8, 11.5)
name = c("Midwest", "South")
qqnorm(midwest, main = "Midwest QQ")
```



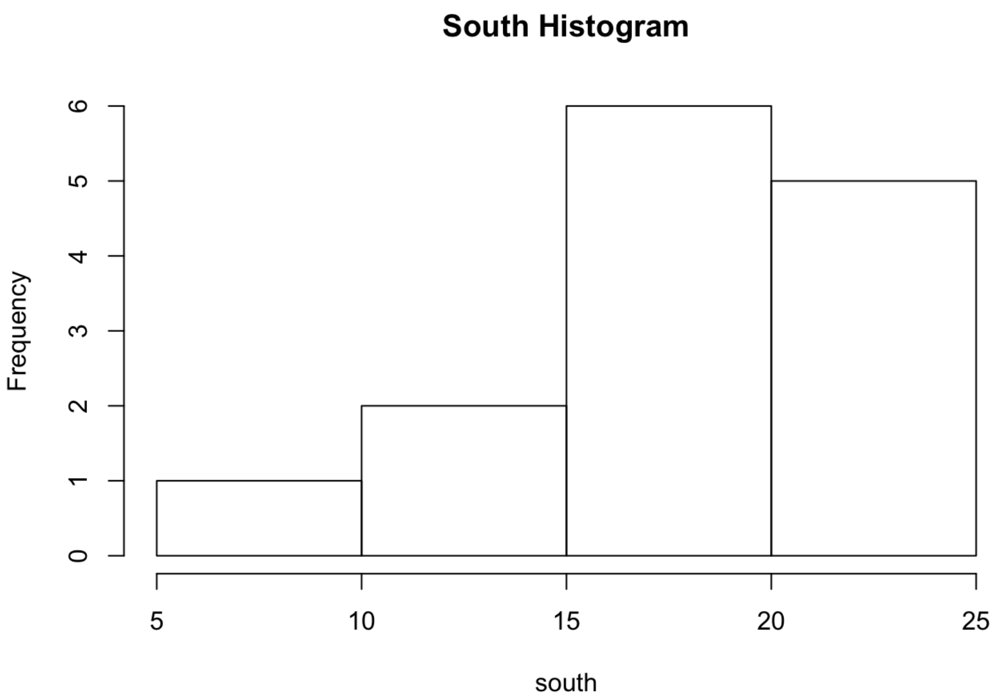

```
qqnorm(south, main = "South QQ")
```



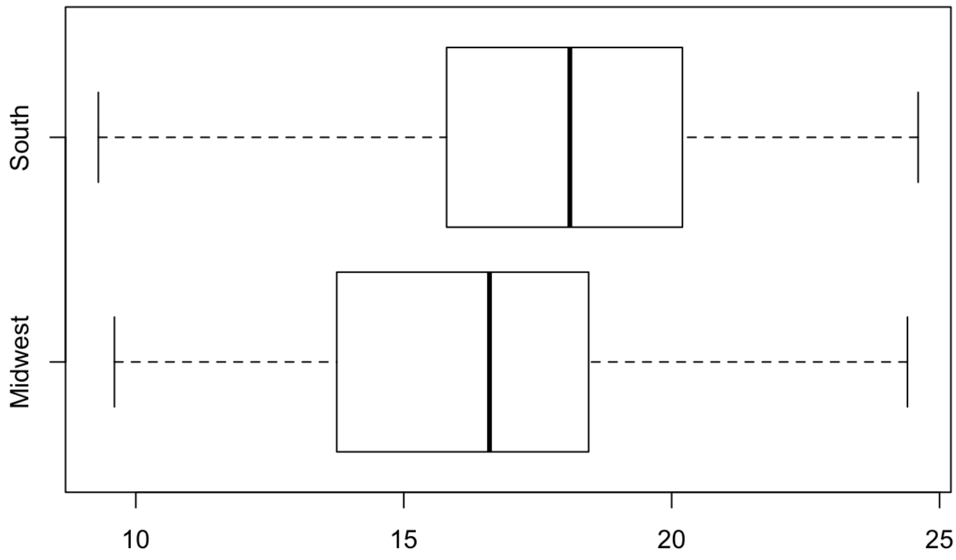

```
hist(midwest, main = "Midwest Histogram")
```




```
hist(south, main = "South Histogram")
```




```
boxplot(midwest, south, horizontal = T, names = name)
```




```
mean(midwest)
```

```
## [1] 16.22667
```

```
mean(south)
```

```
## [1] 17.68571
```

```
sd(midwest)
```

```
## [1] 4.055062
```

```
sd(south)
```

```
## [1] 4.42247
```

The QQ plots and histograms tells us if the data is skewed and normal, the boxplots shows how the data compares.

b. Perform a two population t test for means at the 10% level assuming equal variance and justify why the assumptions of the test are resonably met (or describe what assumptions we are assuming are met). As part of this test, specify your hypotheses, calculate a p value and make a conclusion in the context of the question. Compute the p value just using your calculator and t table. Then find an exact p value by running the same test in R.

Assumptions: Independence, Normality, Equal variance
Ho: $\mu_M = \mu_S$
Ha: μ_M =/= μ_S
S_p = sqrt((((n_M - 1) * s_M^2) + ((n_S - 1) * s_S^2)) / ((n_M + n_S - 2))
S_p = sqrt((((15 - 1) * 4.06^2) + ((14 - 1) * 4.42^2)) / (15 + 14 - 2)) = 4.237153
t = $\bar{X}_M - \bar{X}_S$ / S_p * sqrt(1 / n_M + 1 / n_S)
t = 16.23 - 17.69 / 4.23715 * sqrt(1 / 15 + 1 / 14)
= -0.93
+- $t_{0.10 / 2}$
df = 15 + 14 - 2 = 27
= +-$t_{(27 , 0.05)}$ = +- 1.703
The value of the test statistic t = -0.93 which does not fall in the rejection region. Thus we do not reject Ho. The test results are not significant at 10%.

c. Construct a 90% confidence interval for the true difference in means (by hand and then again in R). Describe how this confidence interval relates to your findings in part b.

$= t_{(27\ ,\ 0.05)} = 1.703$
$(\bar{X}_M - \bar{X}_S)$ +- $t_{a/2} * S_p *$ sqrt$(1 / n_M + 1 / n_S)$
(16.23 - 17.69) +- (1.703) * (4.23715) * sqrt(1 / 15 + 1 / 14)
= (-4.14 , 1.22)
For the two tailed test at 10% level the null hypothesis is not rejected and it is confirmed that the confidence interval (-4.14 , 1.22) contains 0.

```
t.test(midwest, south, var.equal=TRUE, mu=0, alternative = "two.sided", conf.level = 0.90)
```

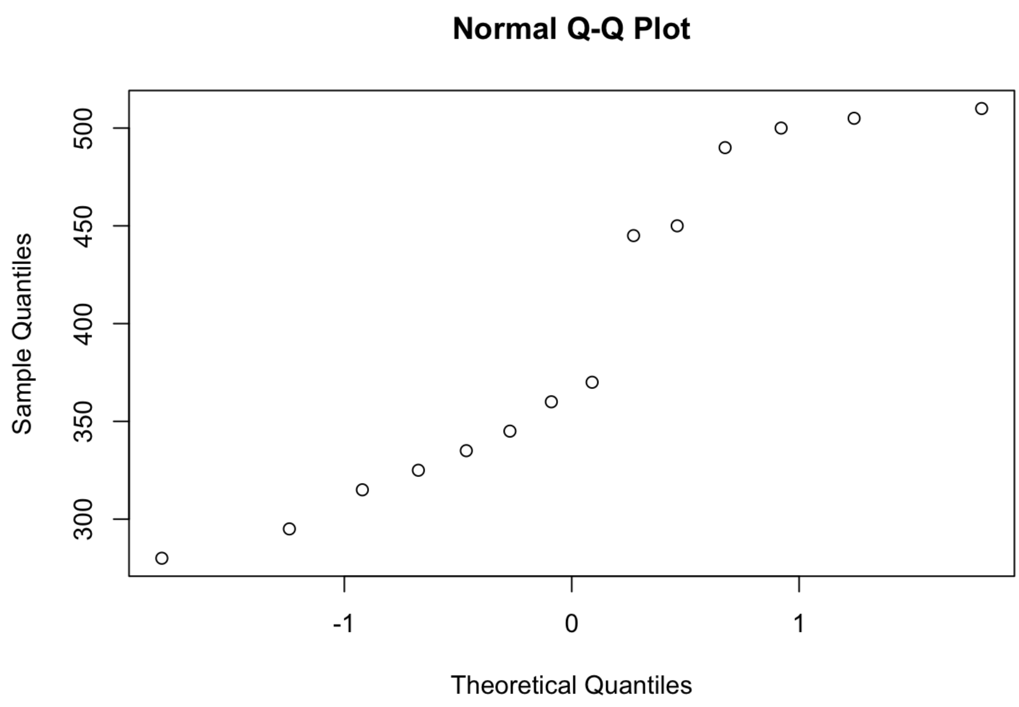
```
##
##  Two Sample t-test
##
## data:  midwest and south
## t = -0.92689, df = 27, p-value = 0.3622
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  -4.140237  1.222142
## sample estimates:
## mean of x mean of y
##  16.22667  17.68571
```

Problem 3

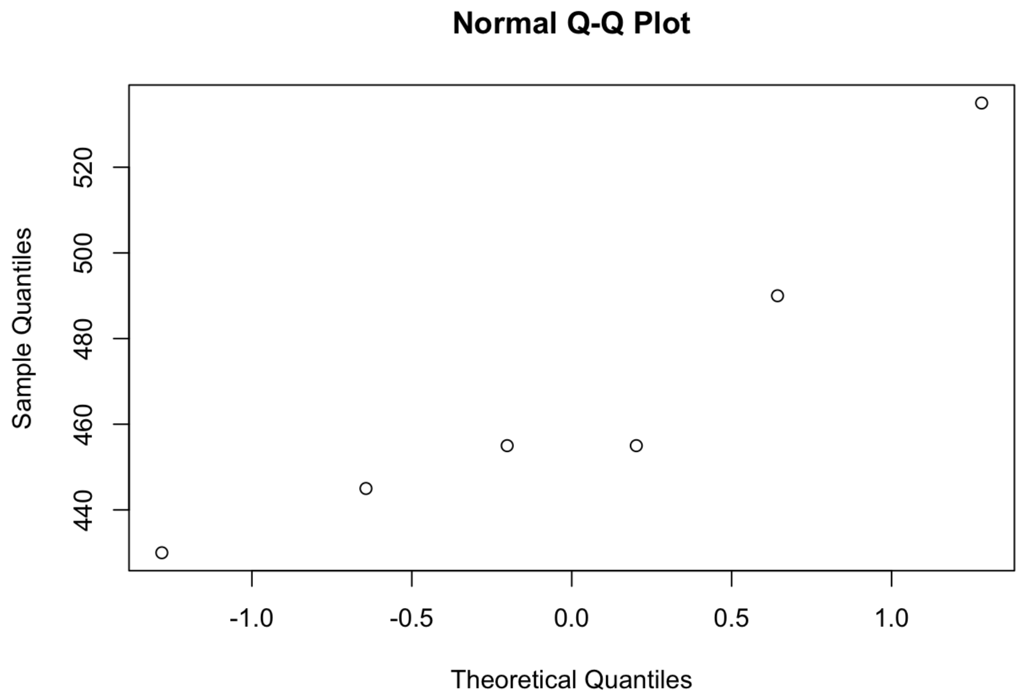
Several neurosurgeons wanted to determine whether a dynamic system (Z-plate) reduced operative time relative to a static system (ALPS plate). The operative times, in minutes, for 14 dynamic replications of the operataion and 6 static replications were obtained and are given below:
Dynamic : 370, 360, 510, 445, 295, 315, 490, 345, 450, 505, 335, 280, 325, 500
Static : 430, 445, 455, 455, 490, 535

a. Graph the data as you see fit. Why did you choose the graph(s) you did and what does it (do they) tell you? Also calculate summary statistics relevant to the research question.

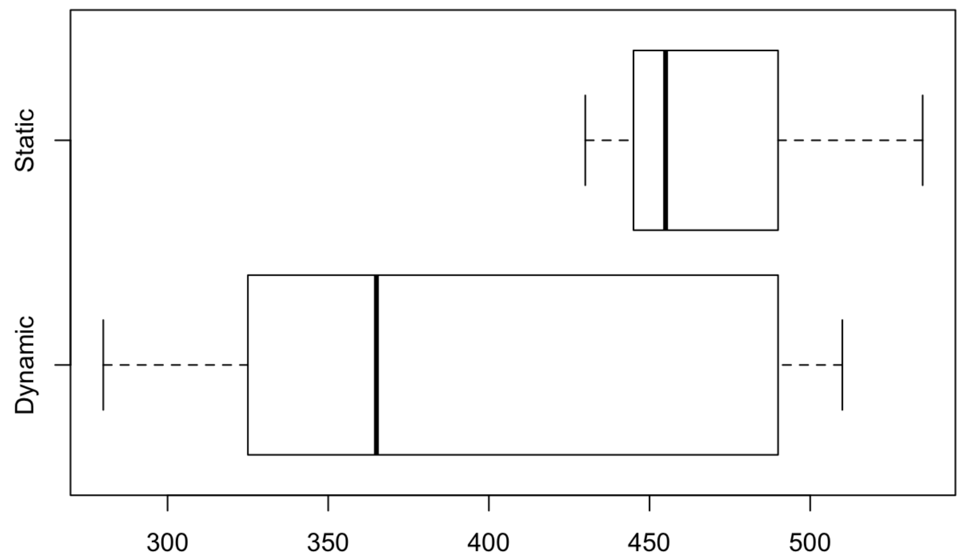
```
Dynamic = c(370, 360, 510, 445, 295, 315, 490, 345, 450, 505, 335, 280, 325, 500)
Static = c(430, 445, 455, 455, 490, 535)
qqnorm(Dynamic)
```




```
qqnorm(Static)
```




```r
boxplot(Dynamic, Static, horizontal = T, names = c("Dynamic", "Static"))
```



We assume SRS (simple random sample) and that the samples are independent. To check large sample size, we refer to the normal probability plots and boxplots. These graphs reveal no outliers and, keeping in mind that the nonpooled t-test is robust to moderate violations of normality, show that we can consider Assumption 3 satisfied. The non-pooled t-test can be usedto carry out the hypothesis test.

b. Perform a two population t test for means at the 5% level not assuming equal variance and justify why the assumptions of the test are resonably met (or describe what assumptions we are assuming are met). As part of this test, specify your hypotheses, calculate a p value and make a conclusion in the context of the question. Compute the p value just using your calculator and t table. Then find an exact p value by running the same test in R.

Ho: $\mu_D = \mu_S$

Ha: $\mu_D < \mu_S$

$t = \bar{X}_D - \bar{X}_S$ / sqrt($S^2_D / n_D + S^2_S / n_S$)

t = (394.6 - 468.3) / sqrt((84.7^2 / 14) + (38.2^2 / 6)) = -2.681

df = ((84.7^2 / 14) + (38.2^2 / 6))^2 / (((84.7^2 / 14)^2 / (14 - 1)) + ((38.2^2 / 6)^2 / (6 - 1))) ~ 17

$t_{(17\ ,\ -2.681)}$

0.005 < pvalue < 0.01
If pvalue <= a, reject Ho, otherwise do not reject Ho.
Because the pvalue is less than the specified significance level of 0.05, we reject Ho. The test results statistically significant at 5% level and provides strong evidence against the null hypothesis.
At the 5% significance level, the data provide sufficient evidence to conclude that the mean operative time is less with the dynamic system than with the static system.

```
t.test(Dynamic, Static, var.equal=TRUE, mu=0, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  Dynamic and Static
## t = -2.0195, df = 18, p-value = 0.02929
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -10.4165
## sample estimates:
## mean of x mean of y
##  394.6429  468.3333
```

c. Construct and interpret an 95% confidence interval for the true difference in means (by hand and then again in R) applying the same assumptions as in part b. Describe how this confidence interval relates to your findings in part b.

= $t_{(17\,,\,0.025)}$ = 2.11
$(\bar{X}_M - \bar{X}_S)$ +- $t_{a/2}$ * S_p * sqrt(1 / n_M + 1 / n_S)
(16.23 - 17.69) +- (2.11) * (4.23715) * sqrt(1 / 15 + 1 / 14)
= (-4.78 , 1.86)
For the two tailed test at 5% level the null hypothesis is not rejected and it is confirmed that the confidence interval (-4.78 , 1.86) contains 0.

```
t.test(Dynamic, Static, var.equal=TRUE, mu=0, alternative = "two.sided", conf.level = 0.95)
```

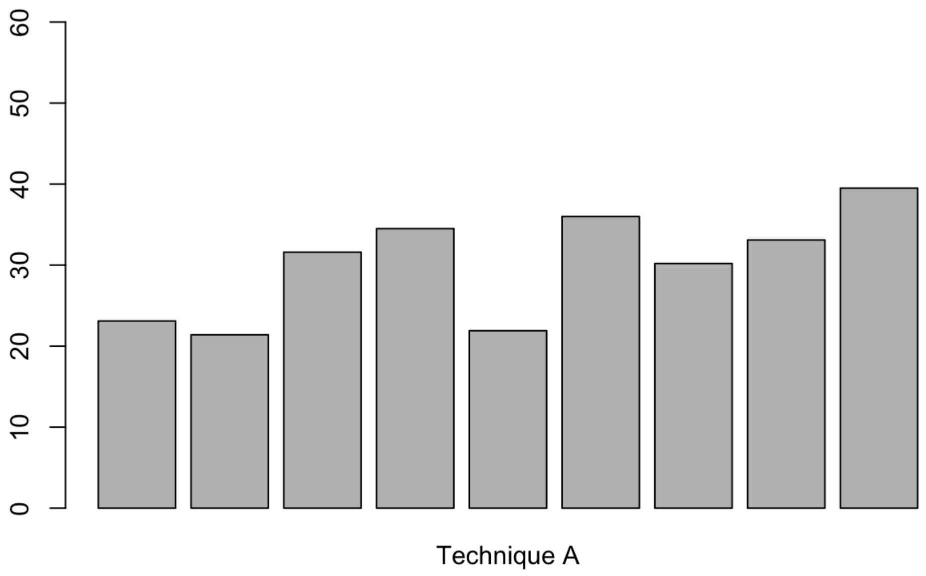
```
##
##  Two Sample t-test
##
## data:  Dynamic and Static
## t = -2.0195, df = 18, p-value = 0.05858
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -150.350695    2.969743
## sample estimates:
## mean of x mean of y
##  394.6429  468.3333
```

Problem 4

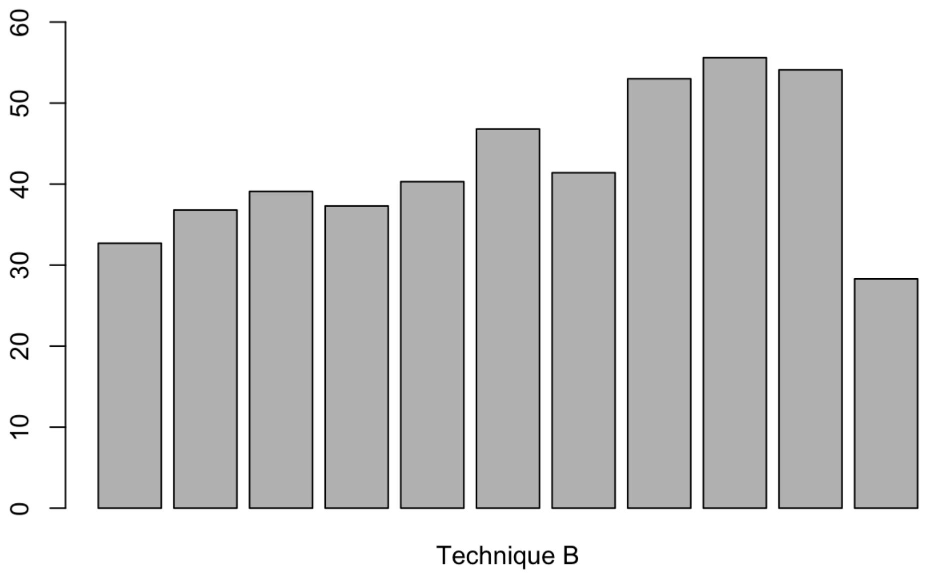
Two new mathematics learning techniques are being tested. Twenty students were randomly selected from a population. nA = 9 of them were randomly assigned to use technique A, and nB = 11 of them were randomly assigned to use technique B. Each student spent 30 minutes learning the technique to which they were assigned, and then were asked to complete a task. The time to complete the task was recorded, in seconds. A shorter time indicates better mastery of the task. The data are below:
TechniqueA : 23.1, 21.4, 31.6, 34.5, 21.9, 36.0, 30.2, 33.1, 39.5
TechniqueB : 32.7, 36.8, 39.1, 37.3, 40.3, 46.8, 41.4, 53.0, 55.6, 54.1, 28.3
We wish to test: Ho : μA − μB = 0 vs. HA : μA − μB =/= 0 using α = 0.05.

a. Graph the data as you see fit. Why did you choose the graph(s) you did and what does it (do they) tell you? Also calculate summary statistics relevant to the research question.

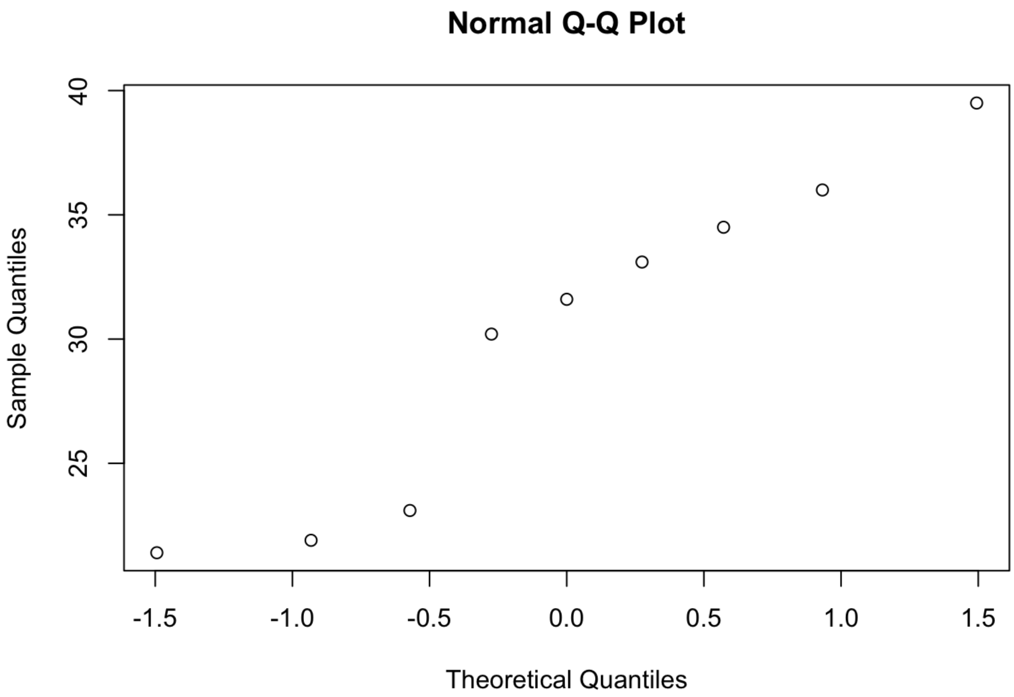
```
A = c(23.1, 21.4, 31.6, 34.5, 21.9, 36.0, 30.2, 33.1, 39.5)
B = c(32.7, 36.8, 39.1, 37.3, 40.3, 46.8, 41.4, 53.0, 55.6, 54.1, 28.3)
barplot(A, ylim = c(0, 60), name = "Technique A")
```



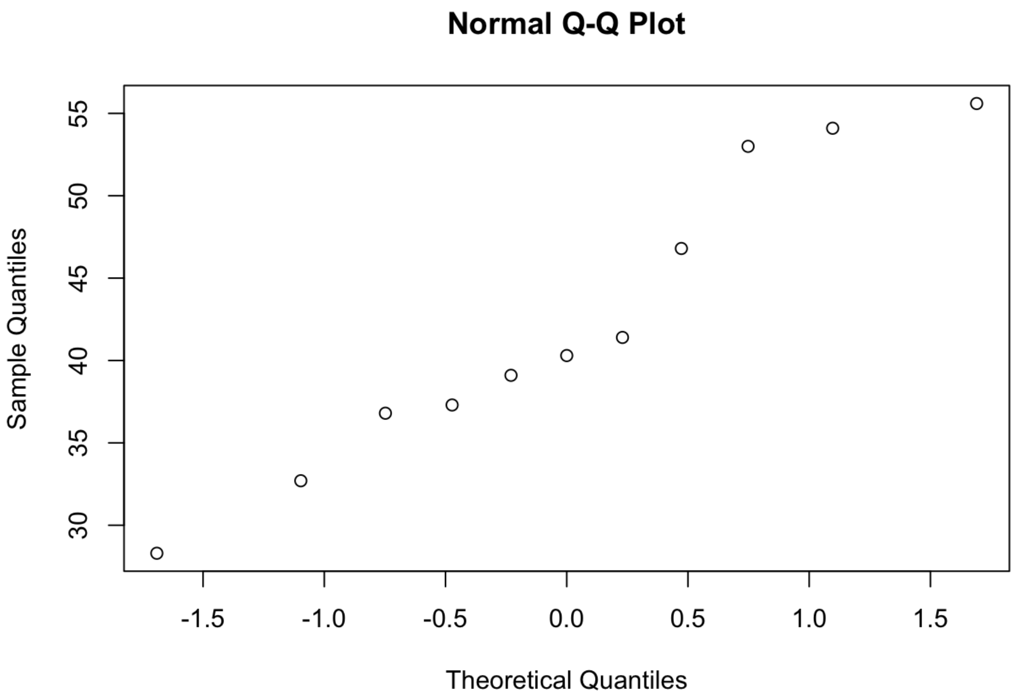

```
barplot(B, ylim = c(0, 60), name = "Technique B")
```



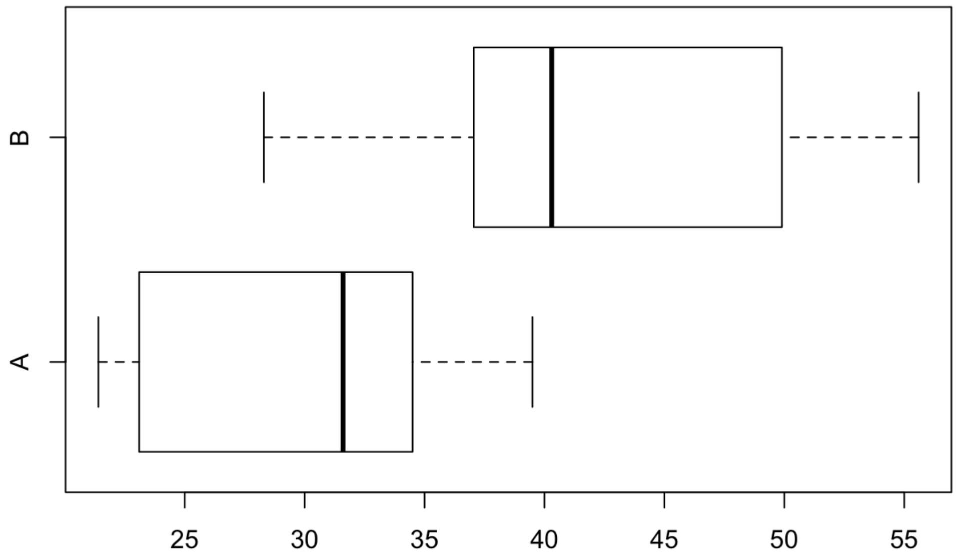

```
qqnorm(A)
```




```
qqnorm(B)
```




```
boxplot(A,B, horizontal = T, names = c("A", "B"))
```




```
mean(A)
```

```
## [1] 30.14444
```

```
mean(B)
```

```
## [1] 42.30909
```

```
sd(A)
```

```
## [1] 6.573643
```

```
sd(B)
```

```
## [1] 8.996272
```

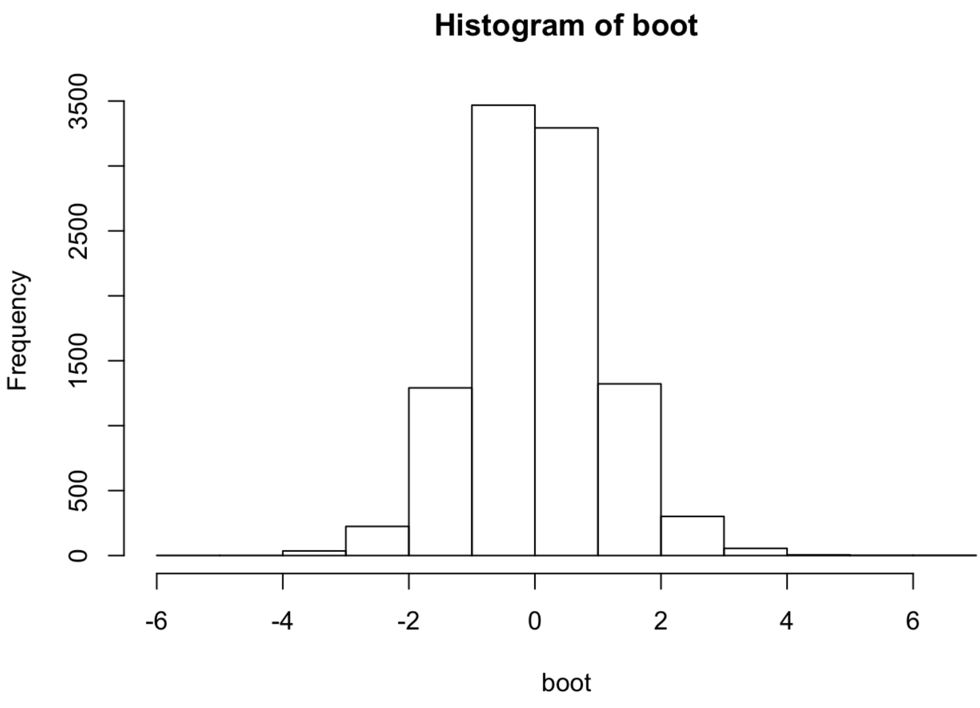
A barplot is used to plot the several groups. It shows that technique B has a higher mean than A.

b. Use the bootstrap to perform the test, using B = 10000 resamplings and set.seed(1). Display the histogram of your generated $\hat{t}$ values. Compute your tobs and a p-value. Make a reject or not reject decision. Finally, state your conclusion in the context of the problem.

Bootstrapping Function:

```r
s.pooled=sqrt(((length(A)-1)*(sd(A))^2+(length(B)-1)*(sd(B))^2)/(length(A)-1+length(B)-1))
se.diff<-sqrt(s.pooled^2/10+s.pooled^2/12)
t.obs<-((mean(A)-mean(B))-0)/se.diff
# Here's one way to do the bootstrap for a difference of two means in R:
# dat1 and dat2 are data from the two groups. nboot is the number of resamples.
#Notice obsdiff is computed as mean(dat1)-mean(dat2) - order matters!
boottwo = function(dat1, dat2, nboot) {
  bootstat = numeric(nboot)
  obsdiff = mean(dat1) - mean(dat2)
  n1 = length(dat1)
  n2 = length(dat2)
  for(i in 1:nboot) {
    samp1 = sample(dat1, size = n1, replace = T)
    samp2 = sample(dat2, size = n2, replace = T)
    bootmean1 = mean(samp1)
    bootmean2 = mean(samp2)
    bootvar1 = var(samp1)
    bootvar2 = var(samp2)
    bootstat[i] = ((bootmean1 - bootmean2) - obsdiff)/sqrt((bootvar1/n1) + (bootvar2/n2))
}
  return(bootstat)
}

B2 = 10000
set.seed(1)
boot = boottwo(A, B, B2)
hist(boot)
```




```r
(t.obs = (mean(A) - mean(B)) /
         sqrt(var(A) / length(A)  + var(B) / length(B)))
```

```
## [1] -3.488601
```

```r
(m.low  = sum(boot < t.obs))
```

```
## [1] 13
```

```r
(m.high = sum(boot > t.obs))
```

```
## [1] 9987
```

```
2 * (p.val = m.low / B2)
```

```
## [1] 0.0026
```

The pvalue < 0.05 so there is strong evidence to reject the null hypothesis, so the means of the two learning technniques are different.

c. Use R to perform a two-group t test for means (i) assuming equal variance and then (ii) not assuming equal variance (Welch’s T) and report the p values from each. Describe how the three p values are related. Explain how the histogram in part (b) and the summary statistics in part (a) hinted at this relationship.

```
t.test(A, B, var.equal=FALSE, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  A and B
## t = -3.4886, df = 17.823, p-value = 0.002654
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.495724  -4.833569
## sample estimates:
## mean of x mean of y
##  30.14444  42.30909
```

The barplots show that almost all of the data in technique B is greater than in A, this shows that the means are going to be different.