HW2

Hanao Li

September 15, 2019

Question 1

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
p_load(MASS, e1071)

blue <- crabs[crabs$sp == "B",]$CL
orange <- crabs[crabs$sp == "O",]$CL

#Parametric Procedure
t.test(blue, orange)
```

```
##
##  Welch Two Sample t-test
##
## data:  blue and orange
## t = -4.2372, df = 197.92, p-value = 3.468e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.000861 -2.189139
## sample estimates:
## mean of x mean of y
##    30.058    34.153
```

```
#Since P value is small, we reject the null hypothesis. There is a significance relationship bet
ween the mean carapace length of blue and orange crabs.


#Nonparametric Procedure
wilcox.test(blue, orange)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  blue and orange
## W = 3378.5, p-value = 7.469e-05
## alternative hypothesis: true location shift is not equal to 0
```

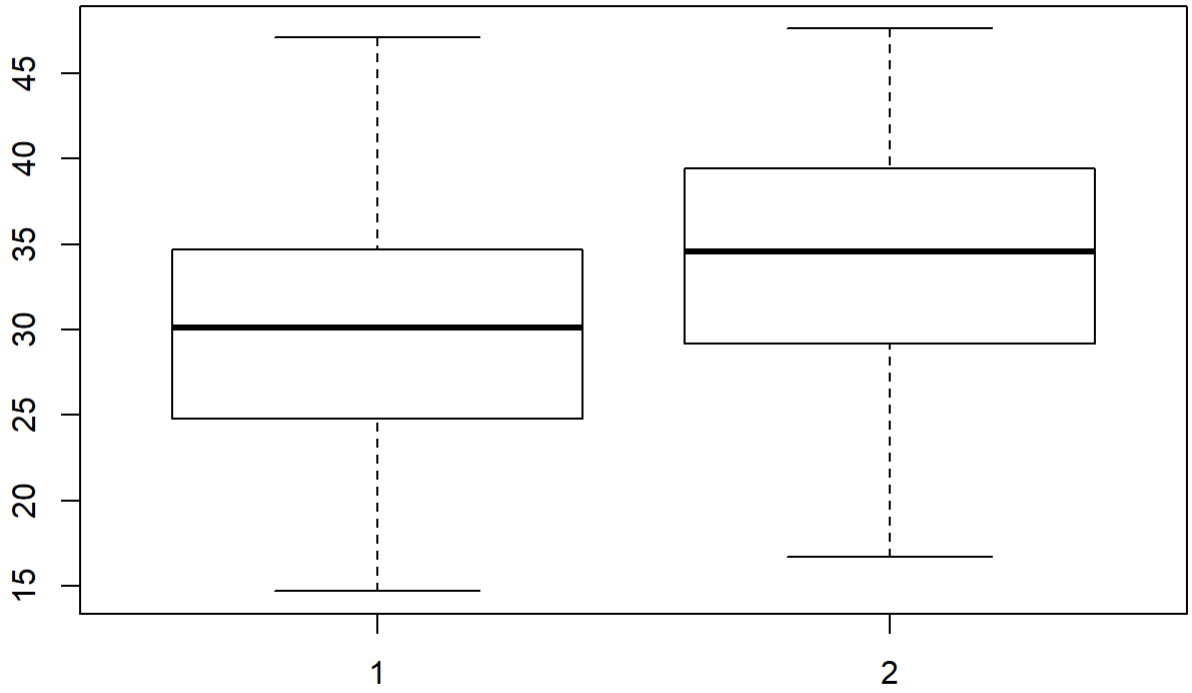
```
#Wilcoxon rank sum test led to the same result of rejecting the null hypothesis.

#Resampling Procedure
z_value <- (mean(blue) - mean(orange)) / sqrt(var(blue) / length(blue) + var(orange) / length(or
ange))
orange_new <- orange + mean(blue) - mean(orange)
z_sample <- rep(NA, 10000)
for (i in 1:10000){
  blue_sample <- sample(blue, length(blue), replace = TRUE)
  orange_sample <- sample(orange_new, length(orange_new), replace = TRUE)
  z_sample[i] <- (mean(blue_sample) - mean(orange_sample)) / sqrt(var(blue_sample) / length(blue
_sample) + var(orange_sample) / length(orange_sample))
}
p_value <- sum(abs(z_sample) >= abs(z_value)) / length(z_sample)
p_value
```

```
## [1] 0
```

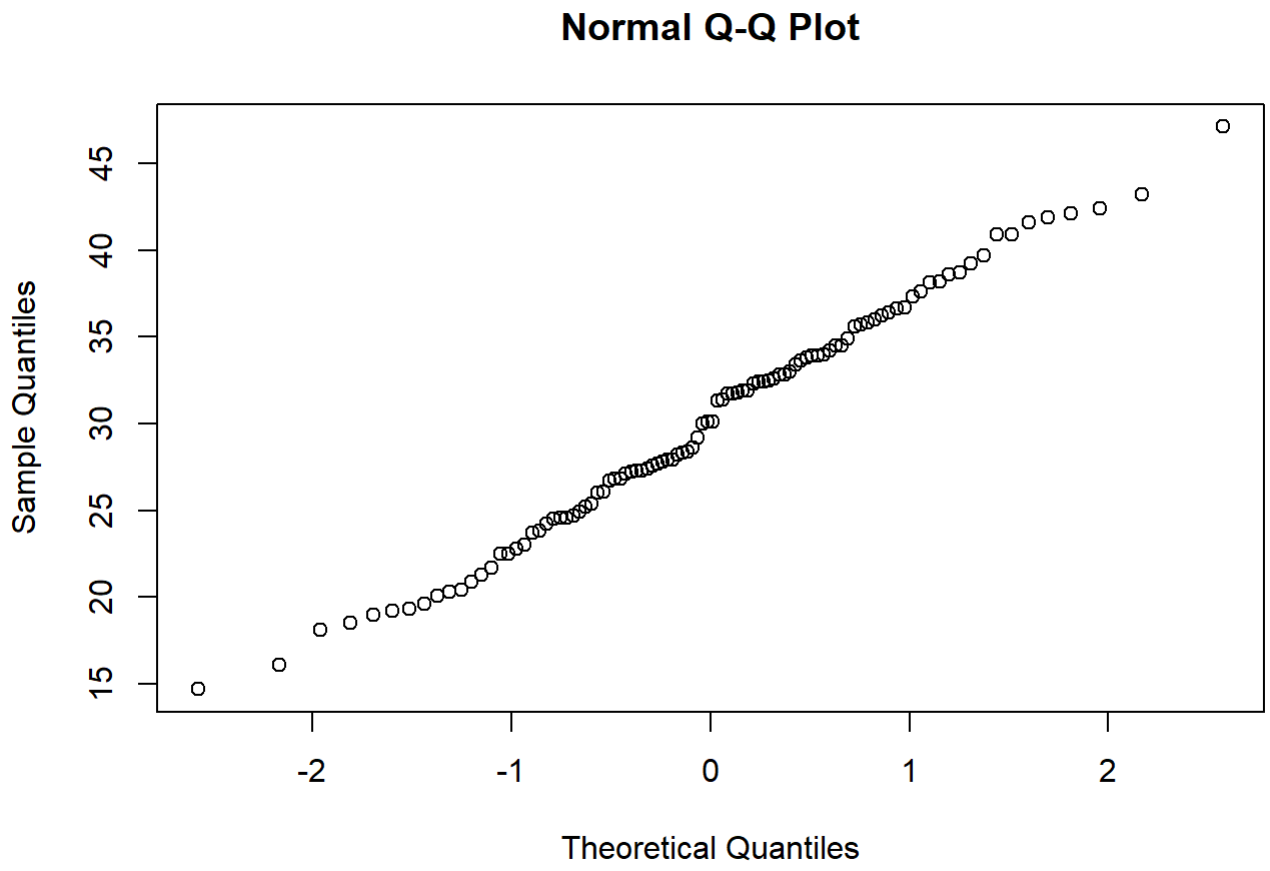
```
#Resampling procesure led to the same result of rejecting the null hypothesis.
```

Question 2

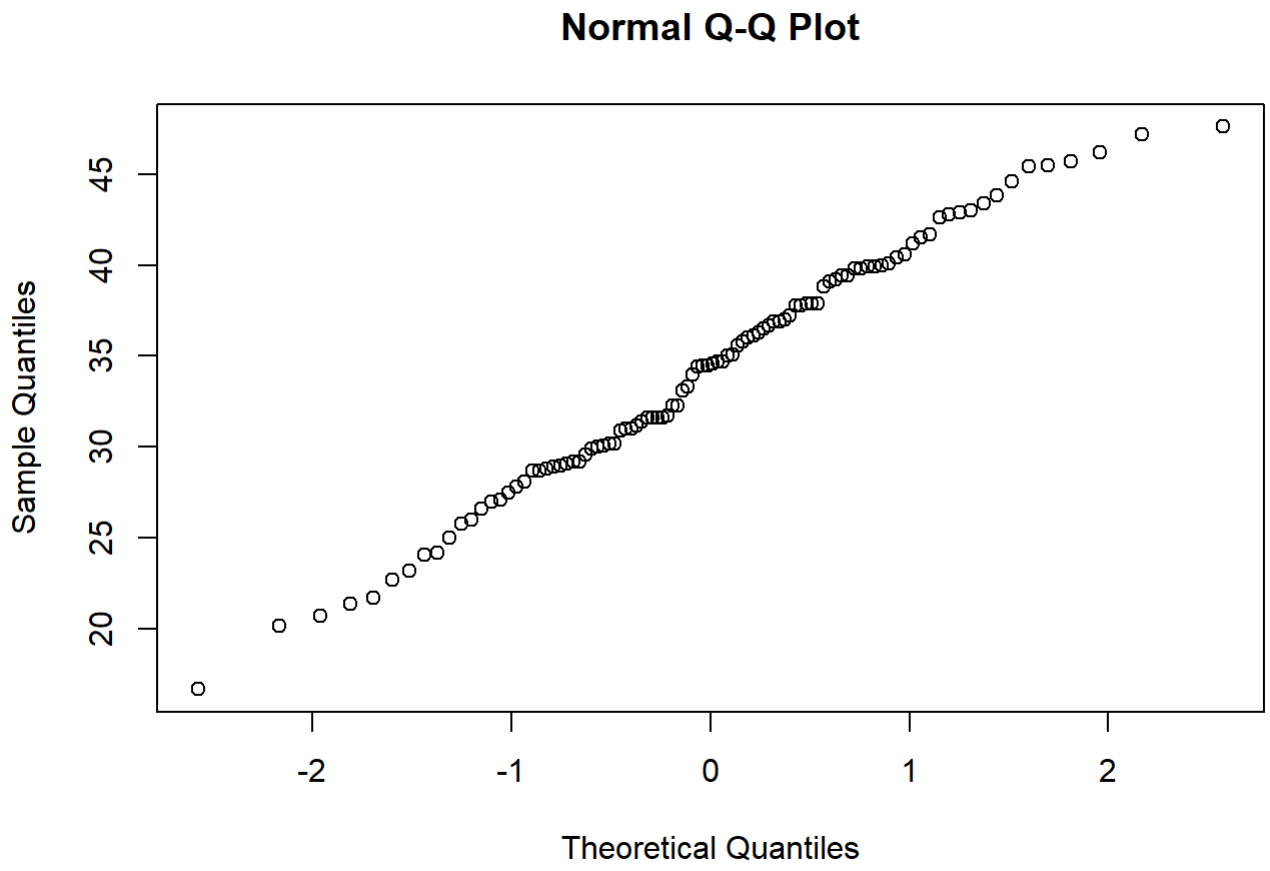
```
#Parametric Procedure

#In the T test, we assume the data satisfy IID normal distribution.

#Check outliers
boxplot(blue, orange)
```




```
#No outliers

#Check normality
qqnorm(blue)
```




```
qqnorm(orange)
```



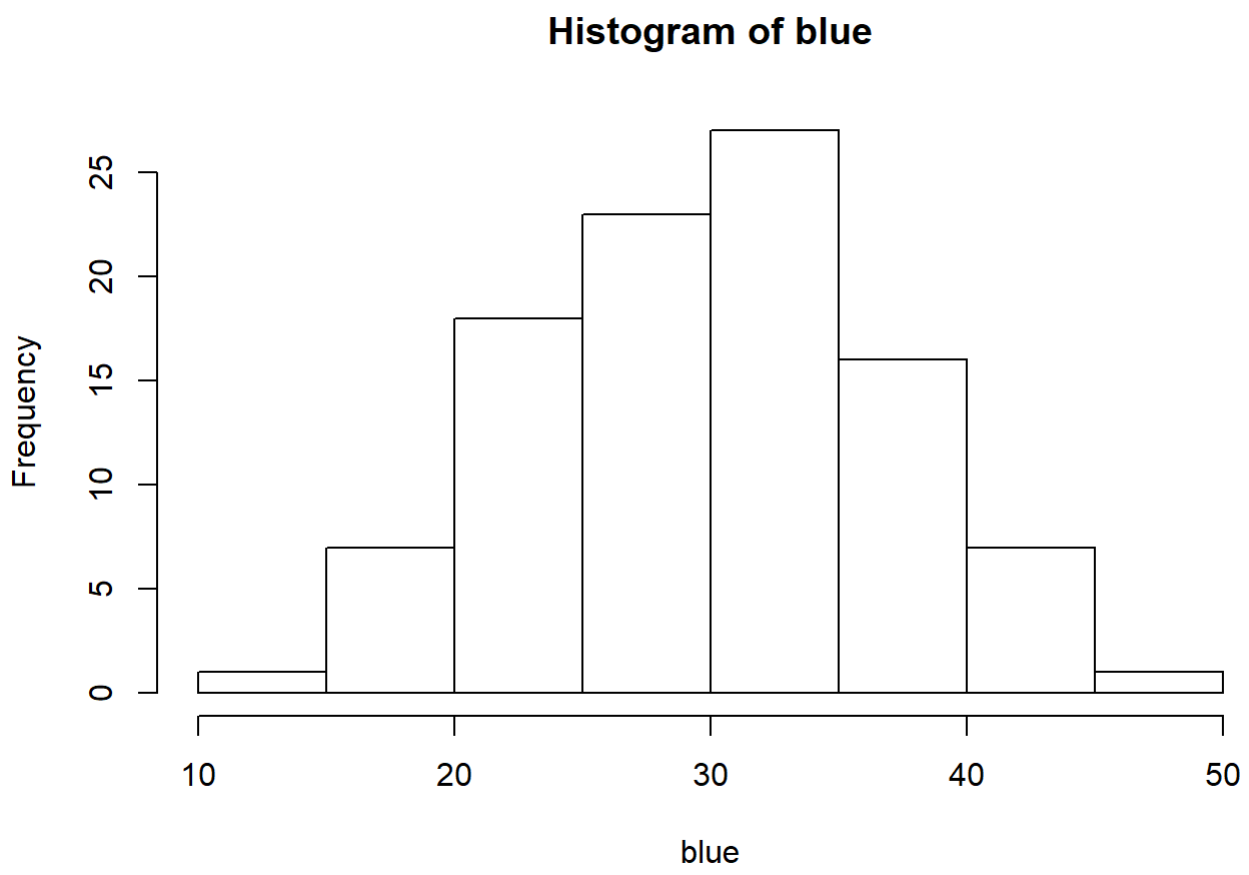

```
#Seems to be normal

#Skewness
skewness(blue)
```

```
## [1] 0.02571518
```

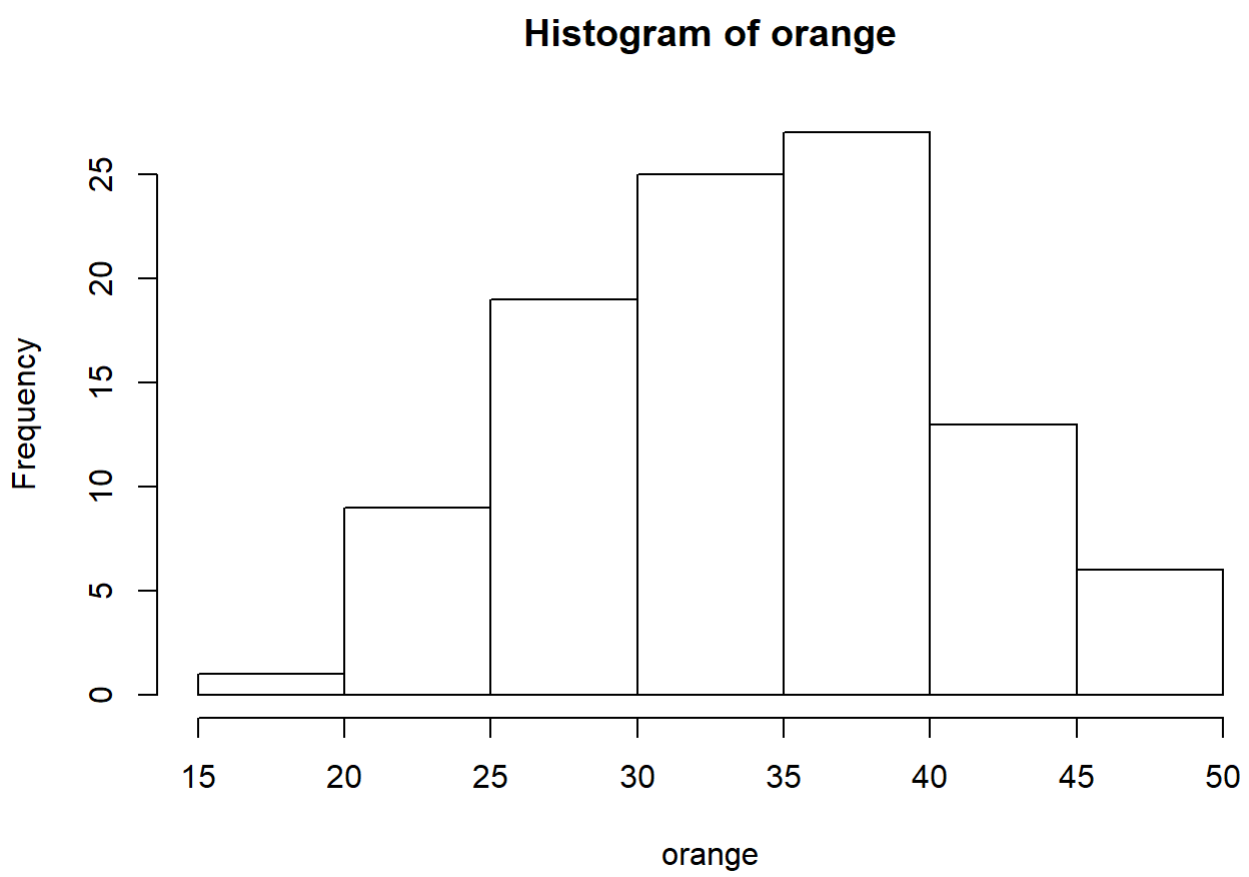
```
skewness(orange)
```

```
## [1] -0.1536776
```

```
#Not very skewed

#Histogram
hist(blue)
```




```
hist(orange)
```




```
#It seems to be normal distribution. So, we conclude it is approximately normal.

#Check Correlation
cor.test(blue, orange)
```

```
##
##  Pearson's product-moment correlation
##
## data:  blue and orange
## t = 22.632, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8777160 0.9429194
## sample estimates:
##       cor
## 0.9161846
```

```
#The blue and orange carbs are not independent. But since the data has no outlier and about to b
e normally distributed, the assumption can be regarded as valid.

#Remedial measures
#Take the log of data, so it can be cloers to normally distributed.

#Nonparametric Procedure
#Nonparametric methods do not make explicit assumptions about underlying distributions and we as
sume they are independent from previous result.

#Bootstrap Procesure
#Bootstrap prcedure assumes the sample should be representative of the population and we assume
 blue and oragne are IID. So it is valid remedial measures will be the same as the parametric pr
ocedure.
```

Question 3

```
#Odds <21
odds1 <- 38 * 61 / (52 * 26)
odds1
```

```
## [1] 1.714497
```

```
se1 <- sqrt(1/38 + 1/52 + 1/26 + 1/61)
CI1 <- exp(log(odds1) + 1.96 * c(-se1, se1))
CI1
```

```
## [1] 0.9213366 3.1904735
```

```
#Odds 21-25
odds2 <- 65 * 153 / (147 * 94)
odds2
```

```
## [1] 0.7197134
```

```
se2 <- sqrt(1/65 + 1/147 + 1/94 + 1/153)
CI2 <- exp(log(odds2) + 1.96 * c(-se2, se2))
CI2
```

```
## [1] 0.4878431 1.0617909
```

```
#Odds >25
odds3 <- 30 * 102 / (42 * 56)
odds3
```

```
## [1] 1.30102
```

```
se3 <- sqrt(1/30 + 1/42 + 1/56 + 1/102)
CI3 <- exp(log(odds3) + 1.96 * c(-se3, se3))
CI3
```

```
## [1] 0.735191 2.302332
```

#From the results, we could see that although for group <21 and group >25, the Odds ratio are larger than 1, their 95% confidence intervals contains 1. So, we can not conclude that there is an association between alcohol consunption and the brease cancer for women with different categories of body mass.

Question 4

```
odd <- (149 * 68) / (129 * 48)
odd
```

```
## [1] 1.636305
```

```
se <- sqrt(1/149 + 1/68 + 1/129 + 1/48)
CI <- exp(log(odd) + 1.96 * c(-se, se))
CI
```

```
## [1] 1.055654 2.536338
```

#Since the 95% confidence interval does not include 1, we can reject the null hypothesis and conclude that there is an association with correct recollection of the orientation between left and right handedness people. The odds ratio is 1.64, so the odds of a correct answer for a left handed person is 1.64 times than a right handed person and there is 95% chance that a left handed person is 1.06 to 2.54 times than a right handed person to get a correct answer.