Name: Jiahong Hu
UNI: jh3561
HW2
STAT 4201

Problem 1

1. A parametric procedure

- Two-sample t test with same unknown but equal variance

  Let X1,...,X100 and Y1,...,Y100 be the CL measurements from orange crabs and blue crabs, respectively.

  H0: Ux – Uy = 0
  H1: Ux – Uy ≠0

```
> t.test(o_cl,b_cl, var.equal=T)

        Two Sample t-test

data:  o_cl and b_cl
t = 4.2372, df = 198, p-value = 3.468e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.189144 6.000856
sample estimates:
mean of x mean of y
   34.153    30.058
```

  P-value = 3.468*10^-5 < 0.05, which shows that sample data provides enough evidence to reject H0 hypothesis and there is a significant difference between blue and orange crabs in CL.

- Welch's modified two-sample t test (No assumption of equal variance of two variables)

```
> t.test(o_cl,b_cl)

        Welch Two Sample t-test

data:  o_cl and b_cl
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:
 2.189139 6.000861
sample estimates:
mean of x mean of y
   34.153    30.058
```

The result under the Welch's test is the same as the previous test in this example.

P-value = 3.468*10^-5 < 0.05, which shows that sample data provides enough evidence to reject H0 hypothesis and there is a significant difference between blue and orange crabs in CL.

2. *A Non-parametric method*

- Wilcoxon Rank-Sum Test

Ho: $F_{orange}(x) = F_{blue}(x) \quad \Rightarrow U_{orange} = U_{blue}$ (Two populations are identical)

Ha: $F_{orange}(x) \neq F_{blue}(x) \Rightarrow U_{orange} \neq U_{blue}$ (Populations I and 2 are shifted from each other)

```
> wilcox.test(o_cl,b_cl)

        Wilcoxon rank sum test with continuity correction

data:  o_cl and b_cl
W = 6621.5, p-value = 7.469e-05
alternative hypothesis: true location shift is not equal to 0

>
```

The *p*-value is 7.469*10^-5. It is smaller than the significance level of 0.05 and it shows that there is enough evidence in the data to reject Ho and suggests that the CL of blue crab and orange crab samples are from different populations.

3. *A re-sampling procedure*

a. Bootstrap Tests

H0: Ux – Uy = 0
H1: Ux – Uy ≠0

```
> b_cl_new<-b_cl-mean(b_cl)
> o_cl_new<-o_cl-mean(o_cl)
> z_obs<-(mean(b_cl)-mean(o_cl))/sqrt((var(b_cl)/100)+(var(o_cl)/100))
>
>
> z<-rep(0,1000)
> for(i in 1:1000)
+ {
+     x <- sample(b_cl_new,replace=T)
+     y <- sample(o_cl_new,replace=T)
+     avg<- mean(x)-mean(y)
+     var_x<-var(x)
+     var_y<-var(y)
+     z[i]<-avg/(sqrt((var_x/100)+(var_y/100)))
+ }
>
> pa<-(sum(abs(z)>=abs(z_obs)))/1000
> pa
[1] 0
```

As we can see above, pa-value is equal to 0, which is less than 0.05. We have enough evidence to reject H0. We can conclude there is significant difference between blue and orange crabs.

Problem 2

1. Parametric method

a. Two-sample t test

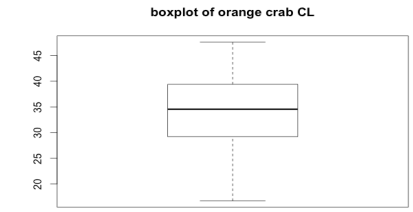
i. Assumptions

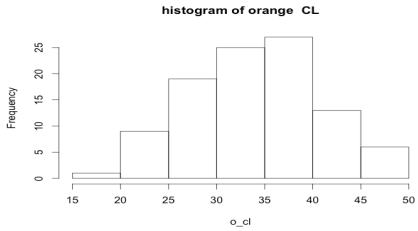
X1, X2… X100 and Y1, Y2, …Y100 should be independent random samples from orange crabs and blue crabs population, respectively.

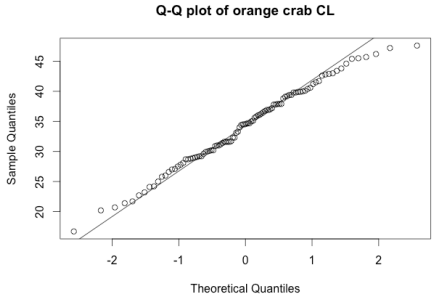
1. Both distributions should be normal.
2. Both distribution should have unknown but equal variance.

ii. Validation

1. Independent
    a. It is true if the orange crabs and blue crabs are randomly selected from their populations, with equal probability being selected.
2. Normal distribution
    a. Orange Crabs

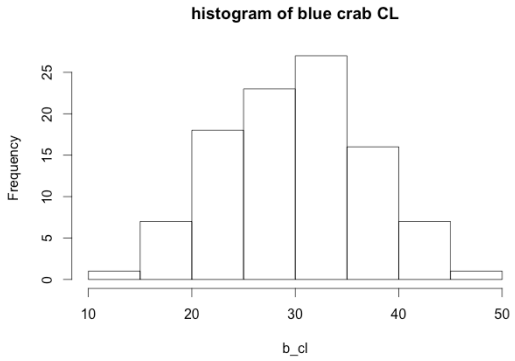


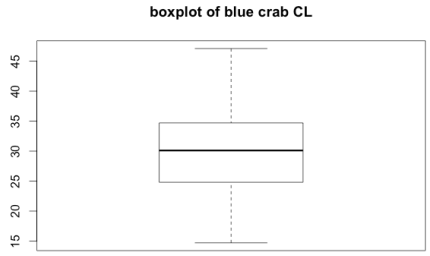


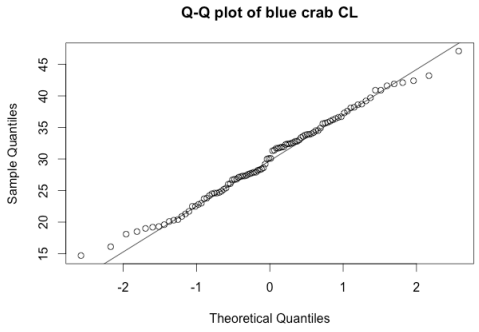


The above graphs show that the orange crab CL is approximately normally distributed.

b. Blue Crab







The above graphs show that the orange crab CL is approximately normally distributed.

3. Equal Variance

    a. Perform F test

```
> var.test(o_cl,b_cl)

        F test to compare two variances

data:  o_cl and b_cl
F = 0.96029, num df = 99, denom df = 99, p-value = 0.8406
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6461232 1.4272157
sample estimates:
ratio of variances
         0.9602902
```

iii. Remedy Measures

1. What if Not normal distribution
    a. Transformation of Data
        i. Taking Logarithm
        ii. Box-cox transformation
        iii. Take more robust, non-parametric approach, which does not need the assumption of normal distribution, such as Wilcoxon Rank-Sun Test
2. What if Variances are not equal
    a. Use Welch's modified two-sample t test.

2. ***Non-Parametric procedure***
    a. Wilcoxon Rank Sum Test
        i. Assumptions
            1. Independent Random Samples taken from two populations whose distributions are identical except one distribution may be shifted to the right or left of the other distribution.
            2. Equal Variance
            3. Continuous Variables
            4. No need for normal distribution

        ii. Validations
            1. Independent - Has been validated in the above " parametric" session already
            2. Equal Variance - Has been validated in the above "parametric" session already
            3. CL is a continues variable
        iii. Remedial Measures
            1. What if not equal variance
                a. If the distributions are normal, we can use parametric method Welch's modified two-sample t test
                b. If the distributions are not normal and do not have equal variances, we can use Kolmogorov-Smirnov Test
3. ***A Re-sampling procedure***
    a. Bootstrap Method
        i. Assumptions
            1. Independent Random Samples
            2. The sample size is large enough to represent the distribution
        ii. Validation
            1. Independent Random Samples
            2. Sample size =100, which is big enough
        iii. Remedial Measures

Problem 3 (19.17)

Alcohol consumption and breast cancer study

	<21 kg/m²		21–25 kg/m²		>25 kg/m²	
	Cases	Controls	Cases	Controls	Cases	Controls
At least one drink per day	38	52	65	147	30	42
Less than one drink per month	26	61	94	153	56	102

Ho: No Association
Ha: There is association

```
> mantelhaen.test(data_1)

        Mantel-Haenszel chi-squared test without continuity correction

data:  data_1
Mantel-Haenszel X-squared = 0.0002009, df = 1, p-value = 0.9887
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.755181 1.329601
sample estimates:
common odds ratio
         1.002043
```

1. **<21kg**

```
> data_2<-matrix(c(38,52,26,61),2,2)
> fisher.test(data_2)

        Fisher's Exact Test for Count Data

data:  data_2
p-value = 0.1174
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8810919 3.3523133
sample estimates:
odds ratio
  1.709275
```

P=0.1174>0.05. No evidence to reject Ho. No association between frequency of drink of alcohol and chance of cancer in the group of people with weight <21kg.

2. 21-25kg

```
> data_3<-matrix(c(65,147,94,153),2,2)
> fisher.test(data_3)

        Fisher's Exact Test for Count Data

data:  data_3
p-value = 0.1154
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.478183 1.081004
sample estimates:
odds ratio
 0.7202373
```

P=0.1154>0.05. No evidence to reject Ho. No association between frequency of drink of alcohol and chance of cancer in the group of people with weight 21-25kg.

3. >25kg

```
> data_4<-matrix(c(30,42,56,102),2,2)
> fisher.test(data_4)

        Fisher's Exact Test for Count Data

data:  data_4
p-value = 0.3811
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7036234 2.3884921
sample estimates:
odds ratio
  1.299486
```

P=0.3811>0.05. No evidence to reject Ho. No association between frequency of drink of alcohol and chance of cancer in the group of people with weight >25kg.

Problem 4 (18.18)

Recollection of comet orientation

	Incorrect	Correct
Left-handed	149	48
Right-handed	129	68

Ho: No Association between right/left handedness and correct recollection of the orientation

Ha: There is association

By Pearson's chi square test

```
> data<-matrix(c(149,48,129,68),2,2)
> chisq.test(data)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 4.4106, df = 1, p-value = 0.03572
```

By Fisher's exact Test

```
> data<-matrix(c(149,48,129,68),2,2)
> fisher.test(data)

        Fisher's Exact Test for Count Data

data:  data
p-value = 0.03547
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.032059 2.602066
sample estimates:
odds ratio
  1.634216
```

The p-value =0.03572 under Chi square test and p-value= 0.03547 under Fisher's Test, which is less than 0.05. It has provided enough evidence to reject Ho and conclude that there is association between right/left handedness and correct recollection of orientation.

How to quantify the association?

Relative Risk = RR =P11/P21= $\frac{\frac{48}{197}}{\frac{68}{197}} = \frac{48}{68} = 0.7059$

Left-handed has negative association with correct recollection of orientation. Correct recollection of orientation is less likely to occur in the left-handed people than in the right-handed people.
See in another approach:

P (LC) = 48/(149+48)=0.2436
P (RC) = 68/(68+129) = 0.3451
n=197
m=197

Hypothesis Test:
Ho: P (LC) = P (RC)
Ha: P (LC) ≠ P (RC)

$$Z = \frac{|0.2436-0.3451|-\frac{1}{2}(\frac{2}{197})}{\sqrt{\frac{2}{197}q_c p_c}} = \frac{0.09642386}{0.04592071} = 2.09979$$

$$q_c = \frac{197*0.2436+197*0.3451}{197*2} = 0.29435$$

```
> x<-c(48,68)
> n<-c(197,197)
> prop.test(x,n)

        2-sample test for equality of proportions with continuity correction

data:  x out of n
X-squared = 4.4106, df = 1, p-value = 0.03572
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.196047850 -0.006997835
sample estimates:
   prop 1    prop 2
0.2436548 0.3451777
```

The p-value is 0.03572 is less than 0.05 and the 95% CI does not include 0. Both show that we have enough evidence to reject H0. There is difference between the correct rate of left hand and right hand people.