HW2

Qiuying Li UNI ql2280

9/21/2017

Problem 1: Determine whether there is a significant difference between blue and orange crabs in mean carapace length (mm) [CL] using each of the following procedures:

$$H_0 : \mu_B - \mu_O = 0 \quad and \quad H_1 : \mu_B - \mu_O \neq 0$$

（B stands for the carapace length of the blue crabs, and O stands for the carapace length of the orange crabs）

a) A parametric procedure

```
## We need a T-test
library("MASS", lib.loc="/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
data = crabs
b_c = data$CL[data$sp == 'B']
o_c = data$CL[data$sp == 'O']
t.test(b_c, o_c)

##
##  Welch Two Sample t-test
##
## data:  b_c and o_c
## 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
```

Conclusion: The p-value is very closed to 0, which means we can reject the Ho, and conclude that there is a significant difference between blue crabs and orange crabs in the mean carapace length.

b) A non-parametric procedure

```
wilcox.test(b_c, o_c)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  b_c and o_c
```

```
## W = 3378.5, p-value = 7.469e-05
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion: The p-value is very closed to 0, which means we can reject the Ho, and conclude that there is a significant difference between blue crabs and orange crabs in the mean carapace length.

c) A re-sampling procedure

```
# I use bootstrap as solution
z_star = (mean(b_c)-mean(o_c))/sqrt(var(b_c)/length(b_c)+var(o_c)/length(o_c)
)
new_val = o_c+mean(b_c)-mean(o_c)
set.seed(123)
z_sample<-rep(NA,1000)
for(i in 1:1000){
  b_sample<-sample(b_c,length(b_c),replace=T)
  o_sample<-sample(o_c,length(o_c),replace=T)
  z_sample[i] = (mean(b_sample)-mean(o_sample))/sqrt(var(b_sample)/length(b_s
ample)+var(o_sample)/length(o_sample))
}
pValue<-sum(abs(z_sample)>=abs(z_star))/length(z_sample); pValue
```

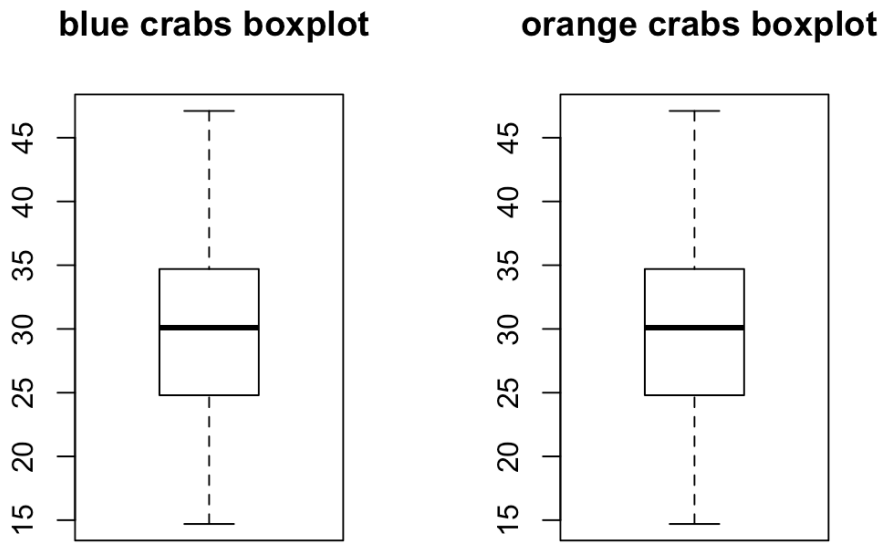
```
## [1] 7.469e-05
```

Conclusion: The p-value is very closed to 0, which means we can reject the Ho, and conclude that there is a significant difference between blue crabs and orange crabs in the mean carapace length.

2. Discuss the assumptions underlying the analyses in (1) above, their validity, and any remedial measures to be taken.

A. Detect an outlier

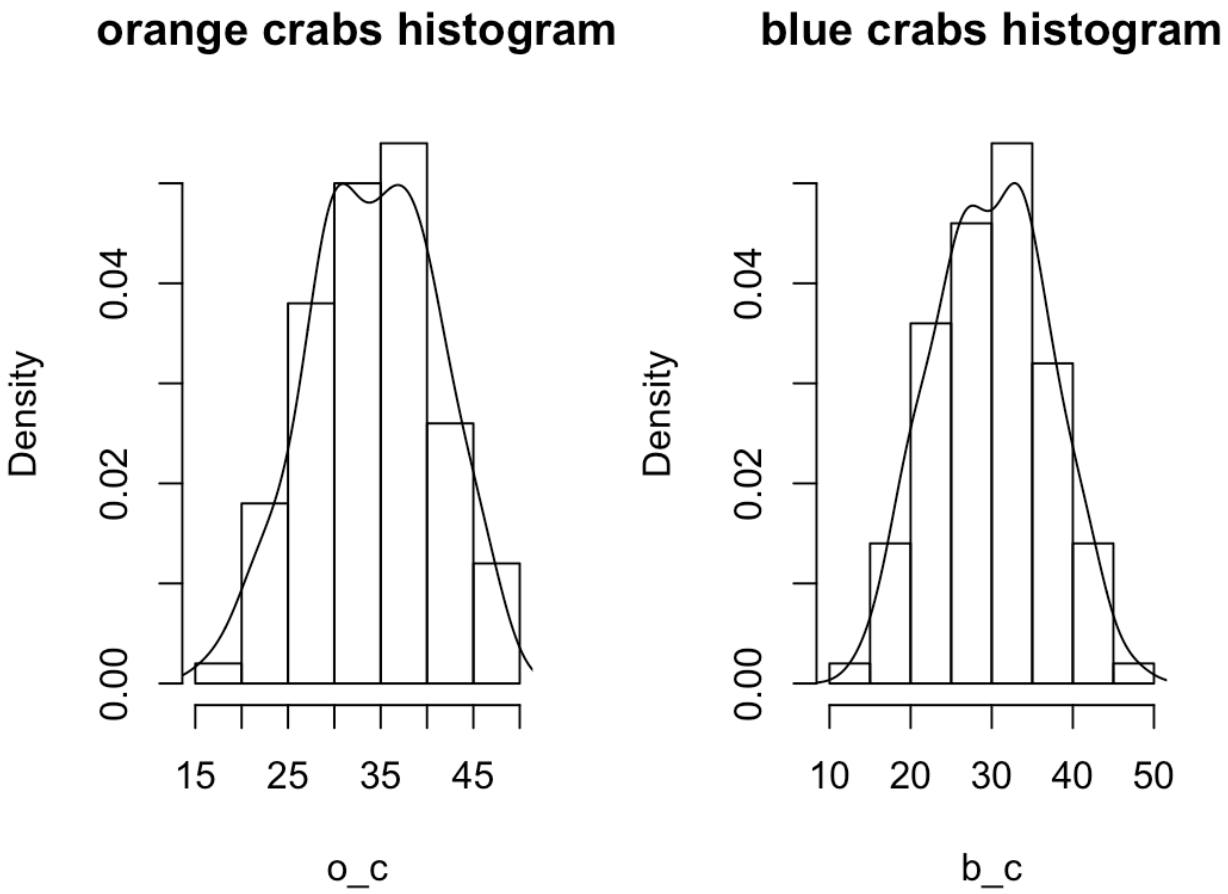
```
par(mfrow = c(1,2))
boxplot(b_c, main = " blue crabs boxplot")
boxplot(b_c, main = " orange crabs boxplot")
```



##Based on the boxplots, there is no outlier detected.

B. Check normality assumption

```
par(mfrow = c(1,2))
hist(o_c, prob = T, main = "orange crabs histogram")
lines(density(o_c))
hist(b_c, prob = T, main = "blue crabs histogram")
lines(density(b_c))
par(mfrow = c(1,2))
```



##From the histograms above, both plots look like bell- shaped symmetric, and the normally assumption held.

```
shapiro.test(b_c)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b_c
## W = 0.99012, p-value = 0.6745
```

```
shapiro.test(o_c)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  o_c
## W = 0.98863, p-value = 0.5557
```

```
library(e1071)
s1<-skewness(b_c);s1
```

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

```
s2<-skewness(o_c);s2
```

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

```
k1<-kurtosis(b_c);k1
```

```
## [1] -0.651835
```

```
k2<-kurtosis(o_c);k2
```

```
## [1] -0.5626413
```

Since the p-value for the Blue Group is p − value = 0.6745, and the p-value for the Orange Group is p − value = 0.5557, so we cannot reject the null hypothesis for both groups. Then we can conclude that the data are approximately normal distributed. Then, check the skewness and kurtosis, the skewness of the two groups are SB = 0.02571518 and SO = −0.1536776. The excess kurtosis of the two groups are KB = −0.651835 and KO = −0.5626413, since both skewness and kurtosis of two groups are close to each other and close to 0. Then the normality is approximately valid.

c. Independence Test

```
cor.test(b_c,o_c)
```

```
##
##  Pearson's product-moment correlation
##
## data:  b_c and o_c
## 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
```

We can conclude that the two-data set are not independent.

p-value is p − value is very closed to 0, then we should reject the null, which means Orange and Blue are not independent.

Summary:

Parametric Test

- **The data of blue crabs and orange data in the carapace length does not have outliers;**
- **Two data set are normally distributed and in bell- shaped symetric**
- **However, the two-data set are correlated, thus we need to do a paired – t test.**

Nonparametric Test Assumption

- **Normal distribution or sample large sample size, at least greater than 30. The sample size is large enough for the central limit theorem to lead to normality of averages.**
  In this problem, we have sample size 100, which is large enough.
- **Two data set are independent**
- **The variances are equal and a scale that is ordinal**

Resampling Test-Bootstrap Method Assumption

- **sub sampling is independent and identical distribution**
- **We have sample size 100, which indicates each sub sampling is independent and identical distribution**
- **we use with replacement method, we just use the empirical distribution to estimate the population.**

Problem 3 Ramsey and Schafer (2nd ed., p. 576), Data Problem 17

```
## <21
exp1<-90*64/177;exp1

## [1] 32.54237

ex1<-38-90*64/177;ex1

## [1] 5.457627

var1<-90*87*64*113/(177*177*176);var1

## [1] 10.26978

z_val1<-ex1/sqrt(var1);z_val1

## [1] 1.703034

pValue1<-1-pnorm(z_val1);pValue1

## [1] 0.04428083

## 21 - 25
exp2<-212*159/459;exp2

## [1] 73.43791

ex2<-65-212*159/459;ex2

## [1] -8.437908

var2<-212*247*159*300/(459*459*458);var2

## [1] 25.88573

z_val2<-ex2/sqrt(var2);z_val2

## [1] -1.658459

pValue2<-1-pnorm(z_val2);pValue2

## [1] 0.9513875

## > 25
ex3<-30-72*86/230;ex3

## [1] 3.078261

var3<-72*158*86*155/(230*230*229);var3

## [1] 12.51782

z_val3<-ex3/sqrt(var3);z_val3

## [1] 0.8700438
```

```
pValue3<-1-pnorm(z_val3);pValue3

## [1] 0.1921382

#calculate total ex
extot<-ex1+ex2+ex3;extot

## [1] 0.09797949

vartot<-var1+var2+var3;vartot

## [1] 48.67332

z_valtot<-extot/sqrt(vartot);z_valtot

## [1] 0.01404396

#calculate odds in <21 group
odds_g1<-38*61/(52*26);odds_g1

## [1] 1.714497

pic_g1<-(38+61)/177;pic_g1

## [1] 0.559322

se0_g1<-sqrt(1/(90*pic_g1*(1-pic_g1))+1/(87*pic_g1*(1-pic_g1)));se0_g1

## [1] 0.3028406

se_g1<-sqrt(1/38+1/52+1/26+1/61);se_g1

## [1] 0.316862

z_val_g1<-log(odds_g1)/se0_g1;z_val_g1

## [1] 1.78021

pValue_g1<-1-pnorm(z_val_g1);pValue_g1

## [1] 0.03752083

CI_g1<-exp(log(odds_g1)+1.96*c(-se_g1,se_g1));CI_g1

## [1] 0.9213366 3.1904735

#calculate odds in 21-25 group
odds_g2<-65*153/(147*94);odds_g2

## [1] 0.7197134

pic_g2<-(65+153)/459;pic_g2

## [1] 0.4749455

se0_g2<-sqrt(1/(212*pic_g2*(1-pic_g2))+1/(247*pic_g2*(1-pic_g2)));se0_g2
```

```
## [1] 0.1874847

se_g2<-sqrt(1/65+1/147+1/94+1/153);se_g2

## [1] 0.1983975

z_val_g2<-log(odds_g2)/se0_g2;z_val_g2

## [1] -1.754288

pValue_g2<-1-pnorm(z_val_g2);pValue_g2

## [1] 0.9603094

CI_g2<-exp(log(odds_g2)+1.96*c(-se_g2,se_g2));CI_g2

## [1] 0.4878431 1.0617909

#calculate odds in >25 group
odds_g3<-30*102/(42*56);odds_g3

## [1] 1.30102

pic_g3<-(30+102)/230;pic_g3

## [1] 0.573913

se0_g3<-sqrt(1/(72*pic_g3*(1-pic_g3))+1/(158*pic_g3*(1-pic_g3)));se0_g3

## [1] 0.2875391

se_g3<-sqrt(1/30+1/42+1/56+1/102);se_g3

## [1] 0.2912111

z_val_g3<-log(odds_g3)/se0_g3;z_val_g3

## [1] 0.9151759

pValue_g3<-1-pnorm(z_val_g3);pValue_g3

## [1] 0.1800496

CI_g3<-exp(log(odds_g3)+1.96*c(-se_g3,se_g3));CI_g3

## [1] 0.735191 2.302332
```

Conclusion:

- For the group of < 21kg/m^2, the odds ratio is 1.714497 and with 95% confidence interval = [1.055654, 2.536338],which means we are 95% confidence that heavy drinkers are 1.055654 to 2.536338 times light drinkers on risk of cancer.
- For group of 21 – 25kg/m^2, the odds ratio is 0.7197134 and with 95% confidence in-

terval = [0.4878431,1.0617909], which means we have 95% confidence that heavy drinkers are 0.4878431 to 1.0617909 times light drinkers on risk of cancer

- For the group of > 25kg/m^2, the odds ratio is 1.30102 and with 95% confidence interval = [0.735191, 2.302332], which means we have 95% confidence that heavy drinkers are 0.735191 to 2.302332 times light drinkers on risk of cancer,
- In short, because the odds ratios are different in 3 groups. The data indicated the trend that heavy drinking is correlated with higher risk of breast cancer among large body mass women and small body mass women. However, women with body mass about 21 – 25kg/m^2 do not show this association.

problem 4

```
logodds<-log(149*68/(129*48));logodds
```

```
## [1] 0.4924406
```

```
pic<-(149+129)/(197+197);pic
```

```
## [1] 0.7055838
```

```
se0<-sqrt(1/(197*pic*(1-pic))+1/(197*pic*(1-pic)));se0
```

```
## [1] 0.2210684
```

```
se<-sqrt(1/149+1/129+1/48+1/68);se
```

```
## [1] 0.2236125
```

```
z_val<-logodds/se0;z_val
```

```
## [1] 2.227548
```

```
pValue<-1-pnorm(z_val);pValue
```

```
## [1] 0.01295532
```

```
CI<-logodds+1.96*c(-se,se);CI
```

```
## [1] 1.55654 2.536338
```

Conclusion:

- Since the p-value is 0.0129, which means that we should reject the null hypothesis at 5% significance level, which means we have enough evidence to show that the data indicate that leftor right-handedness is associated with correct recollection of the orientation. And the 95% confidence interval for the difference of the proportion is [1.55654, 2.536338], which means we have 95% confidence that the odds of a correct answer for a right handed person are approximately 1.636305 times that for a right handed person within the range of [1.55654, 2.536338].