W4201 ADA homework 2 Wenxin Liang wl2455

Consider the data on Chicken Weights by Feed Type (chickwts) in R library MASS

Question 1.

Determine whether there is a significant difference in the mean weights of chicks fed meatmeal vs. those fed casein using each of the following procedures:

The hypothesis for the test we need to do is:

$$H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

a) A parametric procedure

Since when we do the parametric procedure to obtain p-value, the assumptions that parametric procedures required are independence, normality and equal variance.

We run the t-test by

```
>
t.test(chickwts$weight[chickwts$feed=="meatmeal"],chickwts$weight[chickwts
$feed=="casein"],var.equal=T)

	Two Sample t-test

data: chickwts$weight[chickwts$feed == "meatmeal"] and
chickwts$weight[chickwts$feed == "casein"]
t = -1.7294, df = 21, p-value = 0.09842
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -102.801277    9.452792
sample estimates:
mean of x mean of y
 276.9091  323.5833
```

Since p-value equal p-value equal 0.09866, then fail to reject the null hypothesis. For the assumption verification we will do it in the discuss part.

b) A non-parametric procedure

Run Wilcoxon Rank Tests by

```
>
wilcox.test(chickwts$weight[chickwts$feed=="meatmeal"],chickwts$weight[chic
kwts$feed=="casein"])

	Wilcoxon rank sum test

data: chickwts$weight[chickwts$feed == "meatmeal"] and
chickwts$weight[chickwts$feed == "casein"]
```

```
W = 38, p-value = 0.09084
alternative hypothesis: true location shift is not equal to 0
```

Since p-value equal p-value equal to 0.09084, then fail to the null hypothesis.

c) A re-sampling procedure

Run re-sampling test by

```
> chi_meat <- chickwts$weight[chickwts$feed=="meatmeal"]
> chi_casein <- chickwts$weight[chickwts$feed=="casein"]
> chi_casein_bar <- chi_casein+mean(chi_meat)-mean(chi_casein)
> na <- length(chi_meat)
> nb <- length(chi_casein)
> z_obser <-
(mean(chi_meat)-mean(chi_casein))/sqrt(var(chi_meat)/na+var(chi_casein)/nb
)
> z_star    <- c()
> k=1000
> for (i in 1:k) {
+     chi_meat_star     <-sample(chi_meat,na,replace=T)
+     chi_casein_star <- sample(chi_casein_bar,nb,replace=T)
+ z_star[i] <-
(mean(chi_meat_star)-mean(chi_casein_star))/sqrt(var(chi_meat_star)/na+var(c
hi_casein_star)/nb)
+ }
>
> p_value <- sum(abs(z_star)>=abs(z_obser))/k
> p_value
[1] 0.098
```

Since p-value equal p-value equal to 0.098, then fail to the null hypothesis.

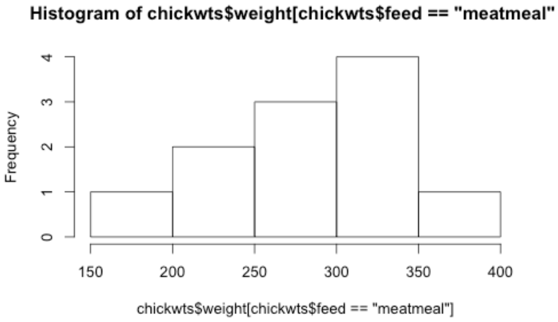
Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken

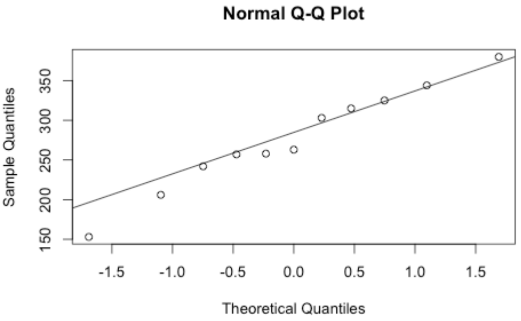
1. For parametric procedures to get p-value, it requires independence, normality and equal variance. When data distribution is skewed, confidence intervals tend to be large, on the average, and p-values may be inflated.

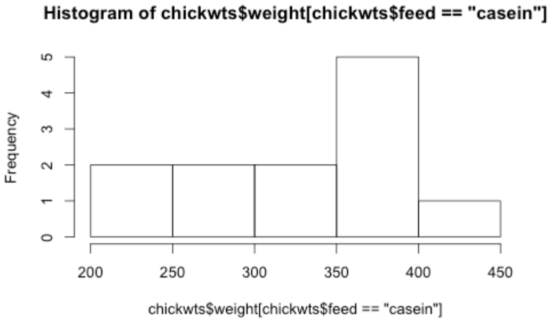
For the independence, based on R we know the two samples do not have equal number of observations then we cannot use correlation to test whether they are independent. In addition based on the description of the data which are the newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types. We still do not know whether they feed in different places or not. So there are still some factors that we lack of information. Then basically the remedial is to consider all the possible factors and make the experiment more rigorous to show the independence of the samples.

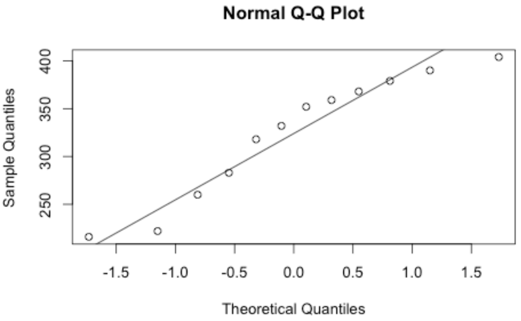
```
> length(chickwts$weight[chickwts$feed=="meatmeal"])
[1] 11
> length(chickwts$weight[chickwts$feed=="casein"])
[1] 12
```

For the normality, we use the histogram and qqplot to show the condition of data,









For both two samples, their histograms don't show good normal density shape. However, look at the qqnorm plots. The points are around the line.

What's more, if we use the Shapiro-wilk test to test the normality we obtain,

```
> shapiro.test(chi_meat)

	Shapiro-Wilk normality test

data:	chi_meat
W = 0.9791, p-value = 0.9612
```

The p-value equals to 0.9612 so we fail to reject the null hypothesis then weights of chicks fed meatmeal follows normal distribution.

```
> shapiro.test(chi_casein)

	Shapiro-Wilk normality test
```

```
data:    chi_casein
W = 0.9166, p-value = 0.2592
```

The p-value equals to 0.2592 so we fail to reject the null hypothesis then weights of chicks fed casein follows normal distribution.

From the graphs and the test we have the opposite conclusion we may consider the reason here can be the insufficient data discrimination. Therefore, the remedial can be collect more data and the distribution of the whole population may be normal. If not we just use the non-parametric procedure to show the conclusion.

Now suppose these two samples have normal distributions. We use F-test to test whether two variances are equal.

```
>
var.test(chickwts$weight[chickwts$feed=="meatmeal"],chickwts$weight[chickwts$feed=="casein"])

    F test to compare two variances

data:  chickwts$weight[chickwts$feed == "meatmeal"] and chickwts$weight[chickwts$feed == "casein"]
F = 1.0145, num df = 10, denom df = 11, p-value = 0.9739
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2877583 3.7182064
sample estimates:
ratio of variances
          1.014541
```

We can see that p-value equals to 0.9739, which is much greater than 0.05. In addition we notice that 1 falls into the 95% confidence interval. Therefore, we fail to reject the null hypothesis then the variances of two samples are equal.

2. For non-parametric procedures do not make explicit assumptions about underlying distributions. Therefore there is no suitable remedial necessary existing.

3. For re-sampling procedure, it involves natural assumptions that the samples are good representatives of the population. Also, the samples should be independent and permutable. Therefore, the remedial can be collect enough data into the samples to make them good representatives of the population, independent and permutable.

Question 2

Using the data for chicks fed soybean and those on sunflower, compute the following, based on a suitable bootstrap method:

a) A 95% confidence interval for the difference in median weight for the two groups

Bootstrap is a non-parametric procedure then there is no assumption of the distribution of the samples. Based on this question, we random replace several times of the data within the samples and get several differences of two medians. Then draw the distribution and calculate the 95% confidence interval.

```
> chicksoy <- chickwts[23:36,1]
> chicksun <- chickwts[37:48,1]
> library(bootstrap)
> mediansoy <- bootstrap(chicksoy,1000,median)
> mediansun <- bootstrap(chicksun,1000,median)
> quantile(mediansoy$thetastar-mediansun$thetastar,c(0.025,0.975))
     2.5%      97.5% 
-131.5000   -48.4875 
```

A 95% confidence interval for the difference in median weight for the two groups is [-131.5000,-48.4875].

b) A 95% bootstrap confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks

```
> varsoy <- bootstrap(chicksoy,1000,var)
> varsun <- bootstrap(chicksun,1000,var)
> quantile(varsoy$thetastar/varsun$thetastar,c(0.025,0.975))
      2.5%      97.5% 
 0.4173211 10.6806409 
```

A 95% bootstrap confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks is [0.4173211,10.6806409]

c) A 95% confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks

```
> var.test(chicksoy,chicksun)

	F test to compare two variances

data:  chicksoy and chicksun
F = 1.2285, num df = 13, denom df = 11, p-value = 0.7412
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3622039 3.9281145
```

```
sample estimates:
ratio of variances
          1.228497
```

A 95% confidence interval for the ratio of the variances of soybean fed to sunflower fed chicks is [0.3622039,3.9281145]

Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken

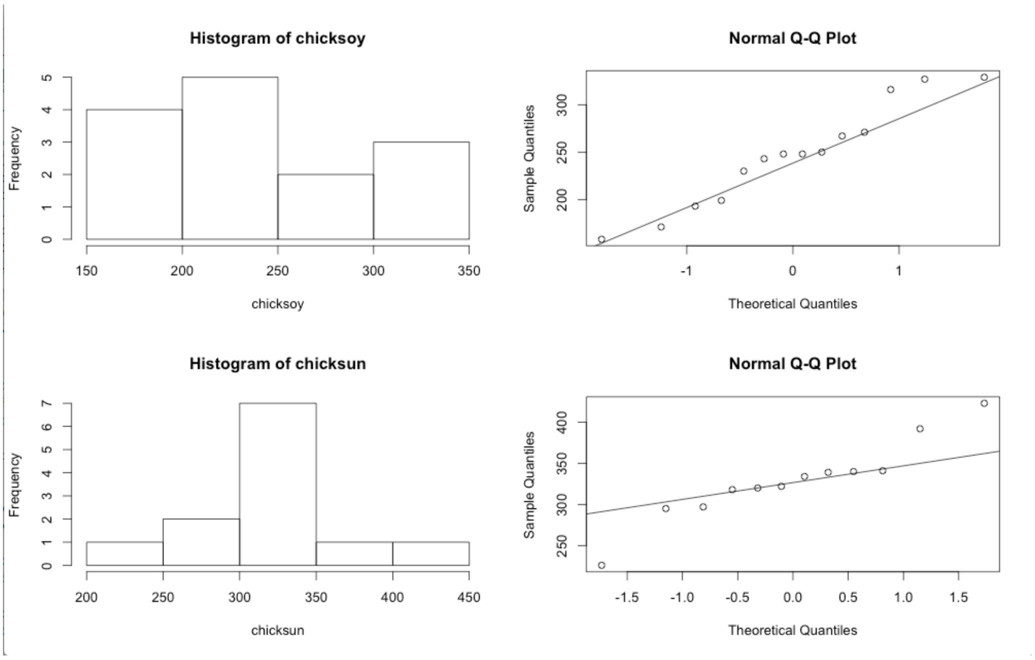
Since part a and part b both using the suitable bootstrap method, part a and part b are using non-parametric procedure which do not make explicit assumptions about underlying distributions.

For part c, it use the parametric procedure to get the 95% confidence interval so the assumptions that independence and normality for the samples are existed.

For the independence, based on R we know the two samples do not have equal number of observations then we cannot use correlation to test whether they are independent. In addition based on the description of the data which are the newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types. We still do not know whether they feed in different places or not. So there are still some factors that we lack of information. Then basically the remedial is to consider all the possible factors and make the experiment more rigorous to show the independence of the samples.

```
> length(chicksoy)
[1] 14
> length(chicksun)
[1] 12
```

For the normality, we use the histogram and qqplot to show the condition of data, for both two samples, their histograms don’t show good normal density shape. However, look at the qqnorm plots. The points are around the line.



What's more, if we use the Shapiro-wilk test to test the normality we obtain,

```
> shapiro.test(chicksoy)

	Shapiro-Wilk normality test

data:	chicksoy
W = 0.9464, p-value = 0.5064
```

The p-value equals to 0.5064 >0.05 so we fail to reject the null hypothesis then weights of chicks fed soybean follows normal distribution.

```
> shapiro.test(chicksun)

	Shapiro-Wilk normality test

data:	chicksun
W = 0.9281, p-value = 0.3603
```

The p-value equals to 0.3603>0.05 so we fail to reject the null hypothesis then weights of chicks fed sunflower follows normal distribution.

From the graphs and the test we obtain the opposite conclusion we may consider the reason here can be the insufficient data discrimination. Therefore, the remedial can be collect more data and the distribution of the whole population may be normal. If not we just use the non-parametric procedure to show the conclusion.

Question 3

Assume that if the weight of a chick is below 258, that chick is classified under "LOW WEIGHT". For chicks fed meatmeal vs. those fed soybean,

a) Determine whether there is a significant difference in the proportions of the chicks classified under "LOW WEIGHT".

For the proportions of the chick classified under "LOW WEIGHT", the hypotheses are,

$$H_0: p_1 = p_2, \qquad H_1: p_1 \neq p_2$$

Based on the R code, we obtain,

```
> meatlow <- chi_meat[which(chi_meat<258)]
> soylow   <- chicksoy[which(chicksoy<258)]
> n_mlow   <-length(meatlow)
> n_slow   <-length(soylow)
> n_m      <-length(chi_meat)
> n_s      <-length(chicksoy)
> x        <-c(n_mlow,n_slow)
> n        <-c(n_m,n_s)
> prop.test(x,n)

    2-sample test for equality of proportions with continuity
    correction

data:   x out of n
X-squared = 0.968, df = 1, p-value = 0.3252
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.7396122   0.1811706
sample estimates:
    prop 1       prop 2
0.3636364 0.6428571
```

Since the p-value is 0.3252>0.05, also from the 95% confidence interval we can conclude that 0 is within the 95% confidence interval, we fail to reject the null hypothesis then the proportions of the chicks classified under "LOW WEIGHT" is significant similar.

b) Construct a 95% confidence interval for the difference in the proportions of the chicks classified under "LOW WEIGHT".

Based on the R code above, we conclude the 95% confidence interval for the difference in the proportions of the chicks classified under "LOW WEIGHT" is [0.3636364, 0.6428571].

Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken

Since there is no warning messages such as " expected counts <5. Chi-square/normal approximation may not be appropriate. in: prop.test(x, n)" happened then the assumption for large n, such that $min(np, nq) \geq$ is satisfied.