W4201 Advanced Data Analysis HW2

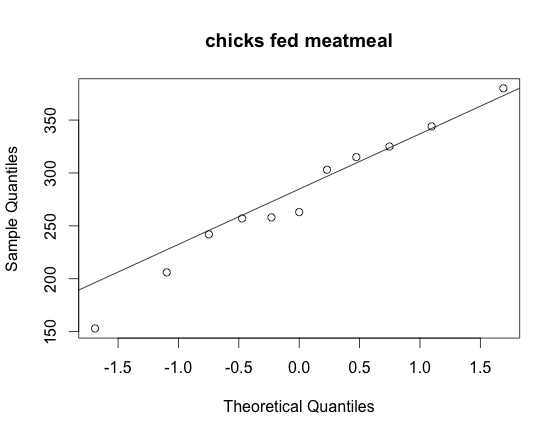
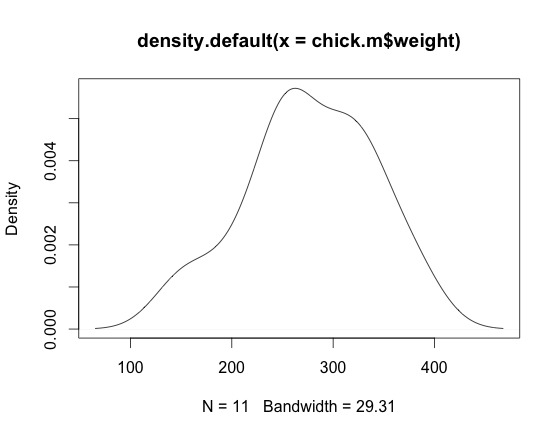
Xi Zhang UNI:xz2364

Q1

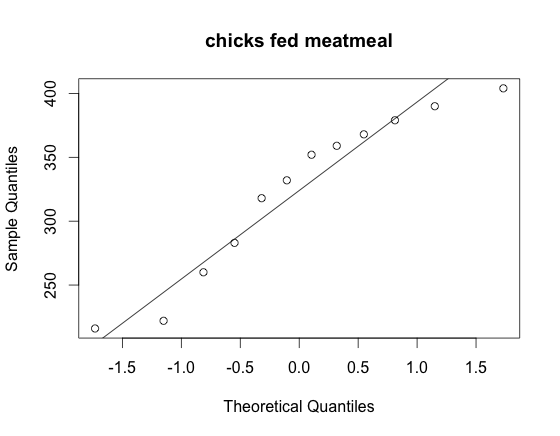
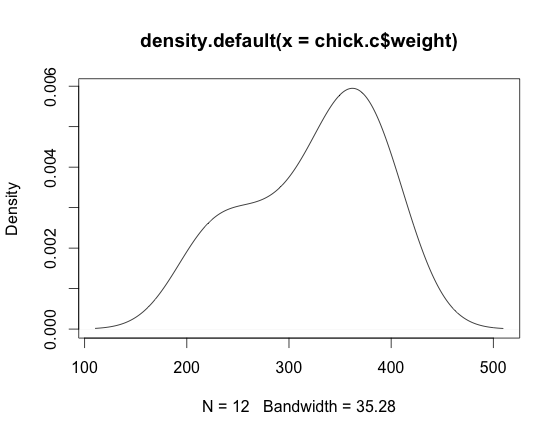
a parametric procedure

In terms of parametric method to test whether there is a significant difference between the mean of two samples, we have the assumption that two distributions are normal and their variances are equal.

Firstly, we use a histogram and qqnorm plot to test the normality.

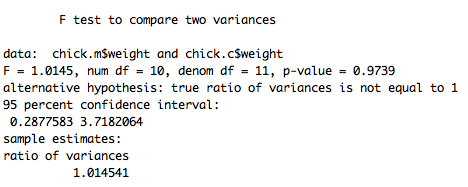


**figure 1 figure 2**



For both two samples, their density plots don’t have good normal density shape. However, look at the qqnorm plots. The points are around the line. I guess the only reason here is the insufficient data discrimination. If possible, we should collect more data and the distribution of the whole population may be normal. We can suppose these two samples have normal distributions for using parametric method here.

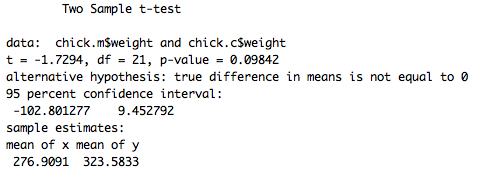
Then, F-test is used here to test whether two variances are equal.



Look at the F-test result. The p-value is much greater than 0.05 and 1 falls into the 95% confidence interval. Therefore, the null hypothesis should not be rejected that is variances of two samples are equal.

Now, let’s use two-sample t test to test the significance of difference of two means. The hypothesis is

The following is the test result.

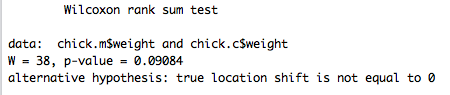


The p-value is 0.09842, which is greater than 0.05. Therefore, the null hypothesis should not be rejected. The means of two samples are significantly similar.

b Non-parametric procedure

We use Wilcoxon rank-sum test here. Assumptions are two samples should be independent and observations in each group should be arranged in order.

The following is the Wilcoxon rank-sum test result.



The p-value is 0.09084, which is greater than 0.05. Therefore, the means of two samples are significantly similar.

c Re-sampling procedure

Referring to randomization distribution of the test statistic mentioned in chapter 1 of The Statistic Sleuth, we firstly combined the two groups into one and re-assign the group several times. Each time, we can calculate the difference of mean between two new groups and finally get several differences. Using these differences, we can draw a distribution to see the 95% confidence interval and calculate the p-value. If the difference of mean of the original two groups falls into the 95% confidence interval and the p-value is greater than 0.05, we can conclude that the difference of mean for two groups is significantly similar.

Let’s re-assign two groups for 1000 times. The 95% confidence interval is [-55.21212,55.95455] and p-value is 0.108. The p-value is greater than 0.05 and the original difference is 46.67424. Therefore, the difference of mean for two groups is significantly similar.

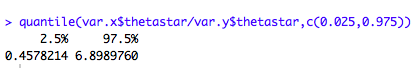
Q2

a Bootstrap is a non-parametric procedure. There is no assumption of the distribution of the dataset. For this question, we just resample each group with replacement several times and get several differences of two medians. Draw the distribution and then calculate the 95% confidence interval.

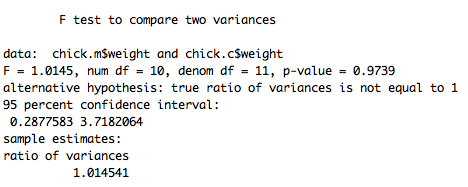
The following is the result.



b In this question, it has the same procedure with I. Instead of difference of two medians, we compare the ratio of two variances. The following is the result.



c In this question, we use F-test here in order to differentiate it with part II.

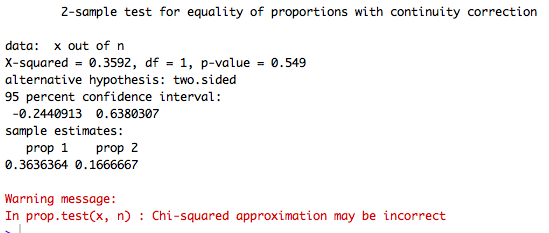


The 95% confidence interval is [0.2877583,3.7182064].

Q3

a & b

We have 11 observations in chicks fed with meatmeal, 4 of them with weight lower than 258 and 12 observations in chicks fed with casein, 2 of them with weight lower than 258. Next, we do the 2-sample test for equality of proportions. When we use Chi-squared test here, we have the assumption that min(np,nq) >= 5 so that the test have the approximate standard normal distribution.



The p-value is 0.549, which is much greater than 0.05 and 0 falls into the confidence interval. Therefore, the difference in the proportion of two samples is significantly similar. The 95% confidence interval is [-0.2440913,0.6380307].

However, there is a warning message, which shows that the assumption isn’t satisfied. We can use fisher test to be the remedial measures, because fisher test does not require the normal distribution and even if the cell expected frequencies is less than 5, we can still use the fisher test.

Assumption validity different place Only use bootstrap for non-parametric

Observation good enough to reflect the original population

Propotional test np>=5, increasing observation number

Quantitative trading

Data

Test trading

Test algorithm

Modeling

Speed Networking10/2

EY on campus interviews

Resume drop deadline(via lionshare) 10/10

On campus interview date:10/21

Behavior based interview on campus

Writing sample –online

Case study interview in house(discussion, write about the case)

Nikhil(Techology)

Operation research projects

Tom undergradte project management data analysis

Eric senior internship interview process

Maduri manager any question

Karan

Strategy critical fee business

Joe Kruse MBA data analysis data sources

Python

SAS

And so on…