Assignment 1

Futong Han,Qinghe Gao,Xinyu Fu

Exercise 1

a) This question is mainly about the two sample t-test and to determine how the sample number and standard deviation influce the power of t-test. Power function is the probability that the t-test rejects the null hypothesis in the given parameters.So the first quenstion:n=m=30;mu=180;sd=5,`nu=seq(175,185,by=0.25)` The H_0 is:mu=nu.

```
#rm(list=ls())
options(digits = 3)
p.value=function(n,m,mu,nu,sd,B=1000){
  p=numeric(B) # p will be an array of realized p-values
  for (b in 1:B) {x=rnorm(n,mu,sd); y=rnorm(m,nu,sd)
                  p[b]=t.test(x,y,var.equal=TRUE)[[3]]}
 return(p)}
```

```
n=m=30;mu=180;sd=5
nulist=seq(175,185,by=0.25);plist=numeric(length(nulist))
d=1
for (i in nulist) { h=p.value(n,m,mu,i,sd); k=mean(h<0.05);plist[d]=k; d=d+1
}
```

b) The we increase the number of sample:n=m=100.The plot shows later

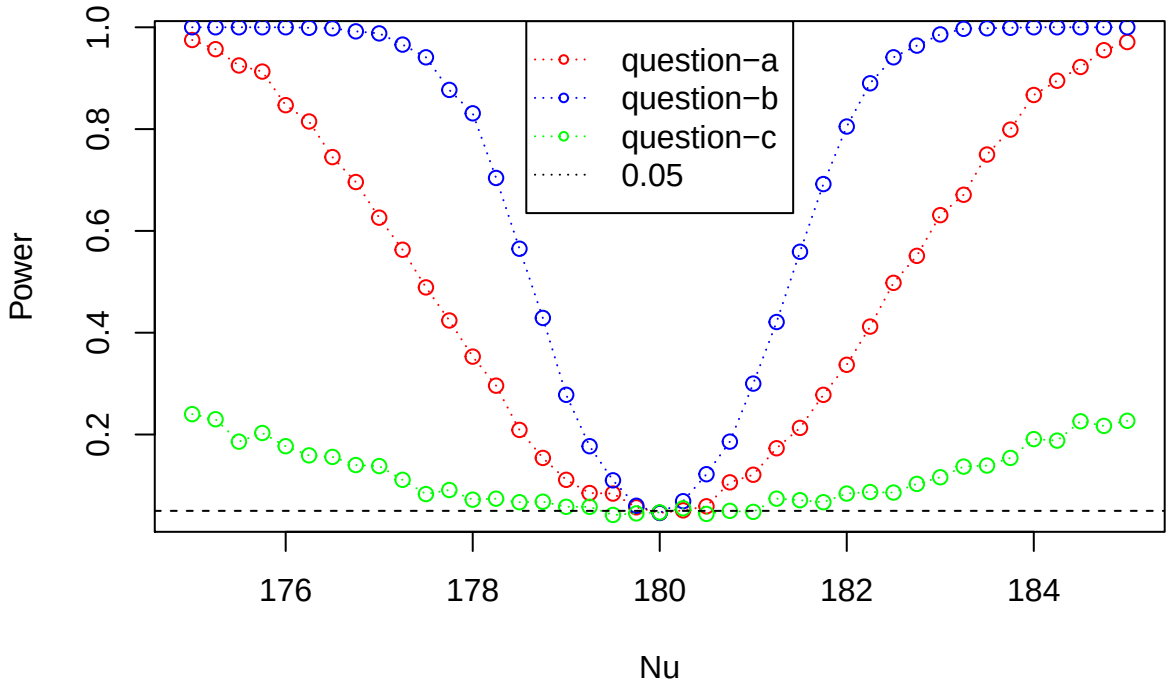
```
n=m=100;mu=180;sd=5
nulist_1=seq(175,185,by=0.25);plist_1=numeric(length(nulist_1))
d=1
for (i in nulist_1) { h=p.value(n,m,mu,i,sd); k=mean(h<0.05);plist_1[d]=k; d=d+1
}
```

c) Finally we increase the standrad deviation.The plot shows later

```
n=m=30;mu=180;sd=15
nulist_2=seq(175,185,by=0.25);plist_2=numeric(length(nulist))
d=1
for (i in nulist_2) { h=p.value(n,m,mu,i,sd); k=mean(h<0.05);plist_2[d]=k; d=d+1
}
```

d) Then we put the three plots toghther:

```
plot(nulist,plist,xlab='Nu',ylab = 'Power',col="red")
lines(nulist,plist,col="red",lty=3)
points(nulist_1,plist_1,col="blue")
lines(nulist_1,plist_1,col="blue",lty=3)
points(nulist_2,plist_2,col="green")
lines(nulist_2,plist_2,col="green",lty=3)
abline(h=0.05,col='black',lty=2)
legend("top",legend=c("question-a","question-b","question-c",'0.05'),pch=c(1,1,1,NA),pt.cex=0.6,col=c("
```



Since this is two sample t-test. And the equation:

$$T = \frac{\bar{x}_n - \bar{y}_m}{S_{x,y}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{1}$$

When compared the result of question a and b sample number has been increased in b. It is clear to see that the plot of b is more sharp and narrow than plot a. Because large sample number increase the significant level of t-test, which decreases the extreme situation. And large sample let the power value become more accurate.

When compared the result of question a and c, standard deviation in c has been increased 15. Larger standard deviation increase the scale of the tail of distribution. Besides larger standard deviation increase the posibility sample the number which are far way from the mean value, which let the even when nu<176 or nu>184 the power value is comparatively larger than a.

Thus, in order to get the accurate the value large sample number and small standrad deviation are necessary.

Exercise 5

a) This is the result of question 5a

```
meatmeal=chickwts[chickwts$feed=='meatmeal',]
sunflower=chickwts[chickwts$feed=='sunflower',]
par(mfrow=c(2,2))
t.test(meatmeal$weight,sunflower$weight)
```

```
##
##  Welch Two Sample t-test
##
## data:  meatmeal$weight and sunflower$weight
## t = -2, df = 19, p-value = 0.04
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -102.57   -1.44
## sample estimates:
```

```
## mean of x mean of y
##       277       329
```

```
wilcox.test(meatmeal$weight,sunflower$weight)
```

```
##
##  Wilcoxon rank sum test
##
## data:  meatmeal$weight and sunflower$weight
## W = 36, p-value = 0.07
## alternative hypothesis: true location shift is not equal to 0
```

```
ks.test(meatmeal$weight,sunflower$weight)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  meatmeal$weight and sunflower$weight
## D = 0.5, p-value = 0.1
## alternative hypothesis: two-sided
```

Category	t-test	Mann-Whitney test	Kolmogorov-Smirnov test
p-value	0.04	0.07	0.1

Two sample t-test assumed that the both two samples were obtained from the normal population and to test whether the mean of two sample are the same. Since the p-value is 0.04, the H_0 is not accepted.

Mann-Whitney test focused on whether the population of two samples are the same and it was based on ranks. We can see the p-value is 0.07, which means H_0 of equal medians is not rejected. The underlying distribution of meatmeal and sunflower are the same.

Kolmogorov-Smirnov test also focused on whether the population of two samples are the same. But it is based on the differences in the histograms. And the p-value 0.1, which we can accept that the weight of meatmeal and sunflower have the same distribution.

b)

```
weightavno_1=lm(weight~feed,data = chickwts)
anova(weightavno_1)
```

```
## Analysis of Variance Table
##
## Response: weight
##           Df Sum Sq Mean Sq F value  Pr(>F)
## feed       5 231129   46226    15.4 5.9e-10 ***
## Residuals 65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(weight~feed,data = chickwts))
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -123.91  -34.41    1.57   38.17  103.09
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     323.58      15.83   20.44  < 2e-16 ***
## feedhorsebean  -163.38      23.49   -6.96  2.1e-09 ***
## feedlinseed    -104.83      22.39   -4.68  1.5e-05 ***
## feedmeatmeal    -46.67      22.90   -2.04  0.04557 *
## feedsoybean     -77.15      21.58   -3.58  0.00067 ***
## feedsunflower     5.33      22.39    0.24  0.81249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.9 on 65 degrees of freedom
## Multiple R-squared:  0.542,  Adjusted R-squared:  0.506
## F-statistic: 15.4 on 5 and 65 DF,  p-value: 5.94e-10
```

So P=$5.9 * 10^{-10}$, it is clear to see that the H_0 can be rejected. So the feed does have different on chick weight.

Besides, the estimated weight for feed6 casein is 259.13. And for horsebean the estimated weight is 323.583-163.383=160.20. For linseed the estimated weight is 323.58-104.83 =218.75. For meatmeal is 323.58-46.67=276.91.For soybean is 323.58-77.15 =246.43. For sunflower is 323.58+5.33 =328.92

Category	casein	horsebean	linseed	meatmeal	soybean	sunflower
Estimated weight	323.58	160.20	218.75	276.91	246.43	328.92

Thus, as for the estimated weight the sunflower is the best feed.

c)

```
meatmeal=chickwts[chickwts$feed=='meatmeal',]
sunflower=chickwts[chickwts$feed=='sunflower',]
horsebean=chickwts[chickwts$feed=='horsebean',]
linseed=chickwts[chickwts$feed=='linseed',]
soybean=chickwts[chickwts$feed=='soybean',]
casein=chickwts[chickwts$feed=='casein',]

shapiro.test(meatmeal$weight);shapiro.test(sunflower$weight);shapiro.test(horsebean$weight);
```

```
##
##  Shapiro-Wilk normality test
##
## data:  meatmeal$weight
## W = 1, p-value = 1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sunflower$weight
## W = 0.9, p-value = 0.4
```

```
##
##  Shapiro-Wilk normality test
##
## data:  horsebean$weight
```

```
## W = 0.9, p-value = 0.5
```

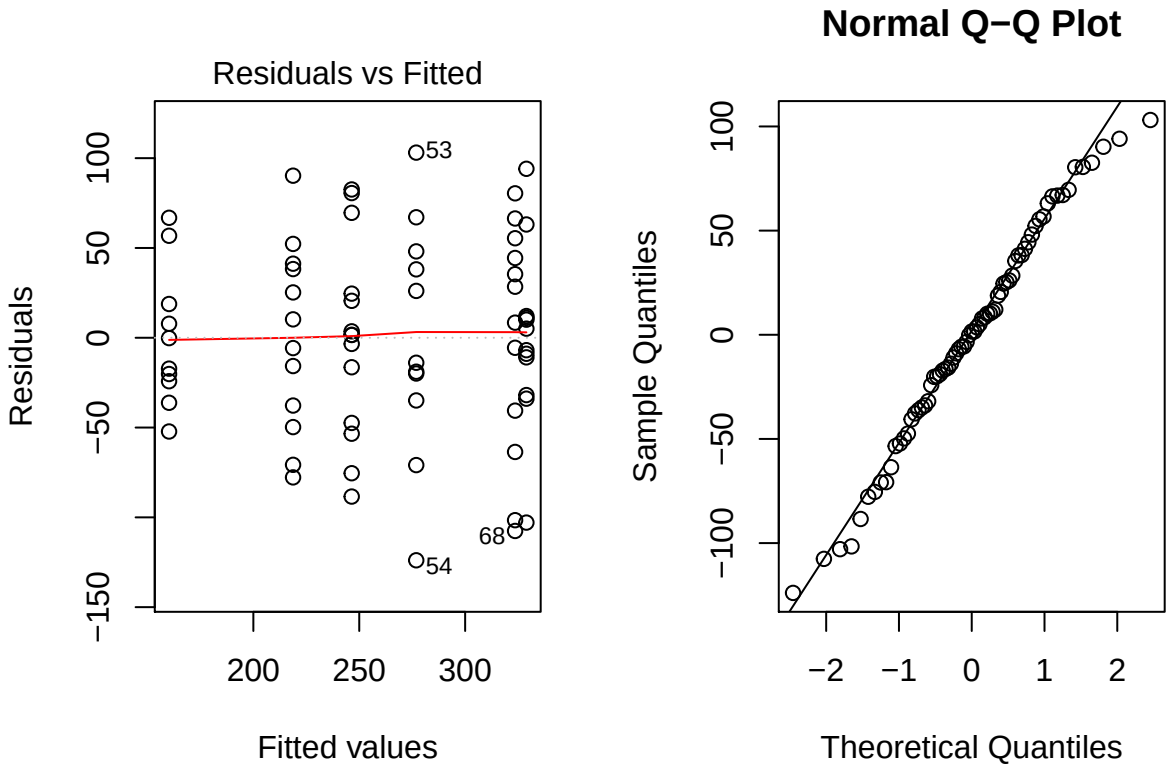
```
shapiro.test(linseed$weight);shapiro.test(soybean$weight);shapiro.test(casein$weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  linseed$weight
## W = 1, p-value = 0.9
```

```
##
##  Shapiro-Wilk normality test
##
## data:  soybean$weight
## W = 0.9, p-value = 0.5
```

```
##
##  Shapiro-Wilk normality test
##
## data:  casein$weight
## W = 0.9, p-value = 0.3
```

```
par(mfrow=c(1,2)); plot(weightavno_1, 1);qqnorm(residuals(weightavno_1));qqline(residuals(weightavno_1))
```




```
#install.packages('car')
#library(car)
#leveneTest(weight~feed,data=chickwts)
```

The first assumption:we need to test of the normality of sample. Since it is really hard to determine the normality by qqlot when the sample number is samll. Thus we use sharipo to test the normality. It is clear to see that all the samples from different feed are normal distribution.

The second assumption: we need to check out the whether the variance of different sample are homogeneous.

We used residuals vs fitted plot and it is clear to see that the there is no relationship between residuals and fitted values, which means we can assume that the variancethe variance of different sample are homogeneous. In order to get the accurate the results whether the variance are homogeneous we used leveneTest to test the variance. It turned out the p-value of leveneTest is 0.59. Then we can certainly get the conclusion that the variance between the samples are the same.

Then we need to determine the normality of residuals. As the qqplot showed we can get the conclusion that the residuals was the normal distribution

Then we can get the conclusion all the assumption of the Anova are stastified.

d)

```
attach(chickwts); kruskal.test(weight,feed)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight and feed
## Kruskal-Wallis chi-squared = 37, df = 5, p-value = 5e-07
```

The difference from the question b is:

When the ANOVA assumptions are not met, Kruskal-Wallis can be used to test whether the samples were from the same population and it is based on the rank. And Kruskal-Wallis actually is a nonparametric alternative to one-way ANOVA. And in this case the p-value is $5 * 10^{-7}$, which the H_0 is rejected and the samples were not from the same population. And the conclusion of b is the sample's mean were not the same.