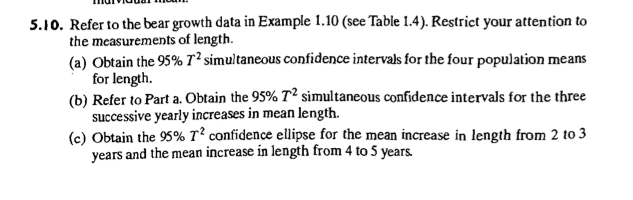
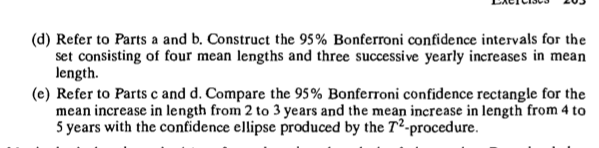
**Problem 5-10:**

****



(a)

From text, section 5.4, the confidence intervals have the following formulas:



Get the results using R code:

[1,] 130.6851 155.8863

[2,] 127.0216 191.5498

[3,] 160.3082 185.9776

[4,] 155.3749 198.9108

So,



(b)

The confidence intervals for the three successive yearly increase in mean length has the following formulas:



Get the results using code

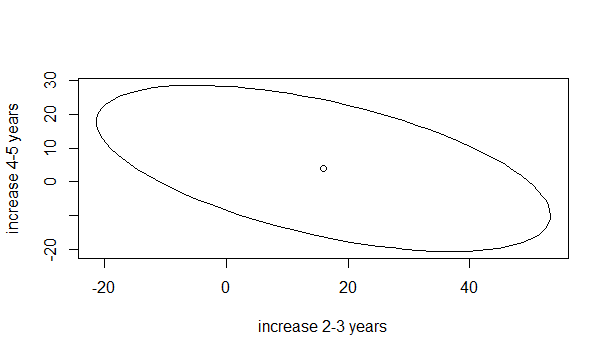
-21.22649 53.22649

-22.73077 50.44505

-20.65385 28.65385



(c)



(d)

The Bonferroni confidence interval for the set consisting of four mean lengths has the following equations:



Get the results using R code:

[1,] 137.3884 149.1831

[2,] 144.1854 174.3860

[3,] 167.1359 179.1498

[4,] 166.9550 187.3307



The Bonferroni confidence interval for the set consisting of four mean lengths has the following equations:



Using R code to get the results:

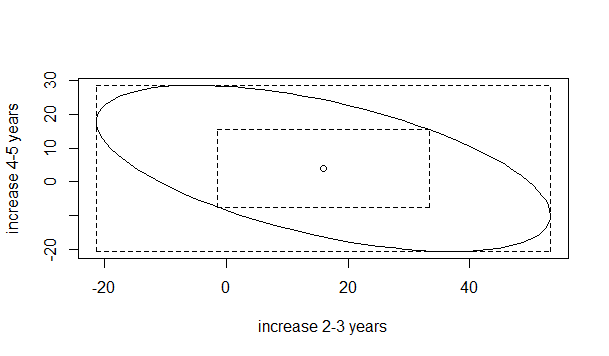
[1,] -1.422784 33.42278

[1,] -3.266772 30.98106

[1,] -7.538521 15.53852

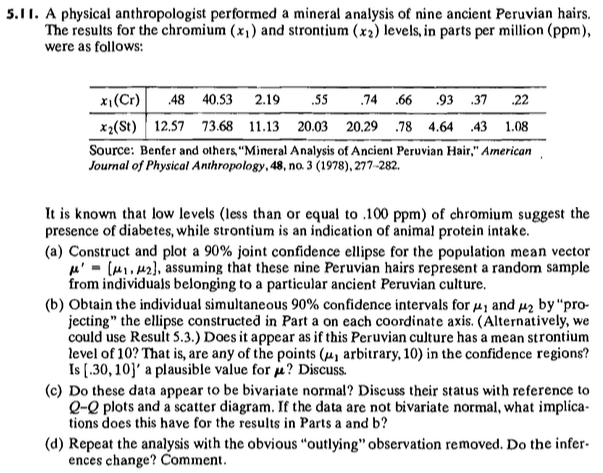


(d)



The confidence rectangle is tangent to the ellipse, which means that the Bonferroni interval falls within the T2-interval. The Bonferroni 95% confidence rectangle is much smaller and more informative than the 95% confidence ellipse.

**Problem 5-11:**



**(a)**

The confidence ellipse will have the form:



p = 2, n = 9, the critical F is 3.2574421, =2.2857, so

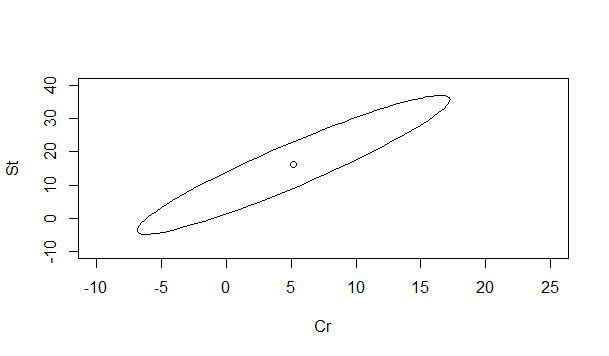


The sample means are { 5.1855556, 16.07} and the sample covariance matrix is



Axes semi-lengths are: 23.870518 and 3.5337114, rho = .942, and the angle of rotation for the

ellipse is 1.06 radians.



(b)

From text, section 5.4, the confidence intervals have the following formulas:



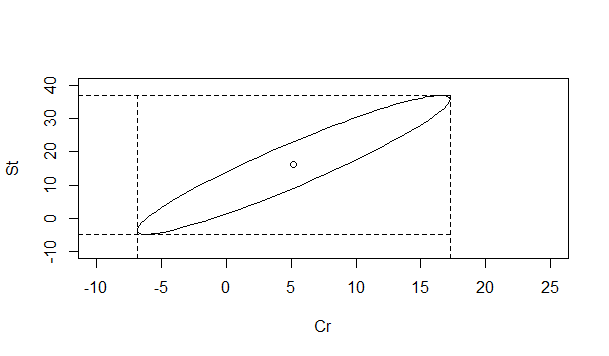
where p, n and S are defined in a and F is the critical F for the indicated degrees of freedom for alpha = .10.

Get the result using R code

[1] -6.881173 17.252284

[1] -4.826957 36.966957





These intervals are the shadows of the ellipse on the two axes. The area of the rectangle they

form is 1008.63. I don't believe the variables can take negative values, and yet the confidence

regions include negative values

**Does it appear that X2 has a mean of 10?**

From the simultaneous confidence intervals, a mean of 10 can't be rejected for some arbitrary values of X1. Also, see the confidence ellipse, which encloses a value of 10 for the possible mean of X2 for many values of X1

**Is [.30, 10] a plausible value?**

The sample mean is { 5.1855556, 16.07}

We can calculate Hotelling's T 2 for this hypothesis.

,which is distributed as a scaled 

The critical F = 3.257442. The scaled value is (8\*2)/7\*F = 7.445581831

Calculate T 2 for this data.

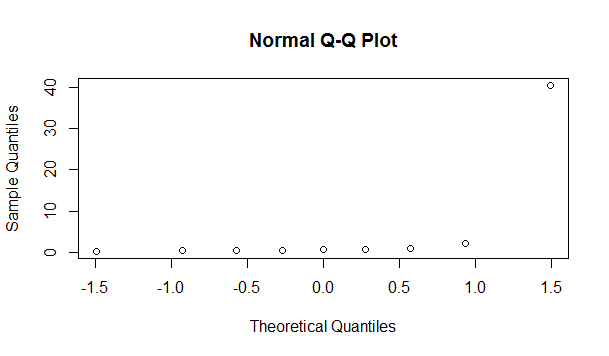
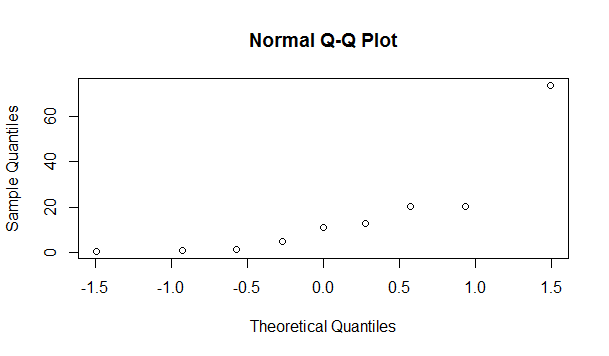
So this is a plausible value. T 2 has a p-value of .497. We can't reject this hypothesis.

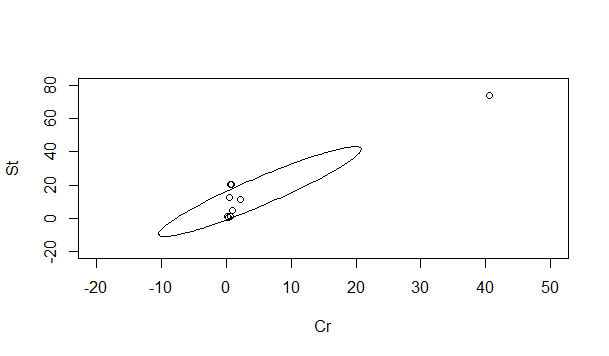
(C)

No. The histograms and Q-Q plots don't appear to show normal data.

Here are the Q-Q plots. The first one is for x1 and second one is for x2.

It is not bivariate normal, which we can tell from the Q-Q plots and scatter plots. So, the results of parta and part b is not reliable, becasue these two methods used in part a and part b has assumptions that the data is normal distributions. However, the data are not normal distributed.





(d)

The second measurement is obviously "wrong." Most likely (looking at the rest of the data) the decimal was off two points but without more information, it's impossible to say if that's what happened or whether the data were recorded in some other units. Dropping this point leaves n =8 with new sample means and covariance matrix below:

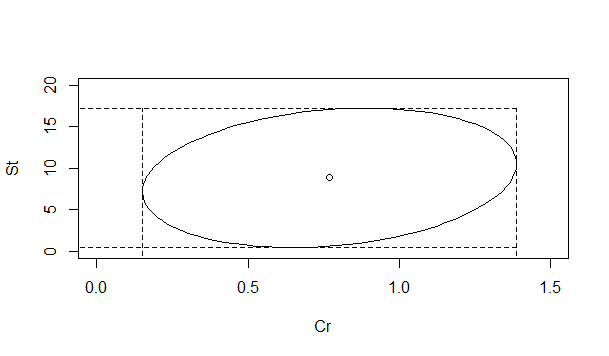
The sample mean is { 0.7675, 8.86875} 

he confidence ellipse will have the form:



p = 2, n = 8, the critical F is 8.08, =2.3333, so

 (Sketch is below)



The ellipse will be bounded by the rectangle formed by these simultaneous confidence intervals. These intervals no longer include negative numbers.

0.1491156 ≤ u1 ≤ 1.3858844

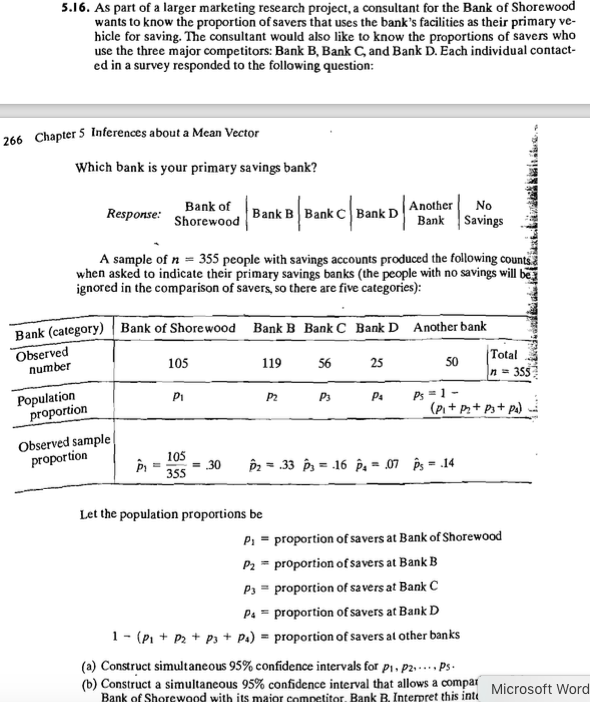
0.468306 ≤ u2 ≤17.269194

The axes half-lengths are 8.4013623 and 0.605781. (Other information about the ellipse was found as before.)

We still wouldn't reject the hypothesis that the mean strontium level is 10 as the value falls in the confidence interval and also, for some values of chromium also in the ellipse (which I generated in SAS; axes have different scales). We would not reject the hypothesis that u = {.3, 10} based on T 2 = 5.3078722 compared to a critical value of 8.08.

the inferences change the rectangular regrion from [0.1516156, 0.469556]'to [1.3883844,17.270444]', which is smaller than the rectangluar region before.

**Problem 5-16:**



**(a)**

We are asked to construct simultaneous confidence intervals for proportions using some banking data. N = 355, so large sample properties can be used.

The proportions are

The covariance matrix can be constructed from the proportions, using σkk = pk\*(1-pk) and σik = -pipk



We have the general result for the confidence interval: 

For this problem, each a is a vector with 1 element equal to 1, the rest zero. So the a'Sa term

just returns the on-diagonal elements (variances) of the covariance matrix.

In this case, q = 4 (one less than the number of proportions because the proportions must sum to

1) The critical chi-square value = 9.488, so we have:

0.3 ± .0749163

0.33 ± .0768708

0.16 ± 0.0499331

0.07 ± .0417117

0.14 ± 0.0567257

(b)

Compare Shorewood with its major competitor. Look at the difference in proportions.



We have the general result for the confidence interval: 

N=355, 

The confidence interval is 0.03 ± 0.1297 (-.0997 to 0.1597). This interval contains 0, so there is

no evidence of a difference. This is a fairly wide confidence interval for a difference in

proportions. From the sample, the point estimate of the difference is -.03. Although these two

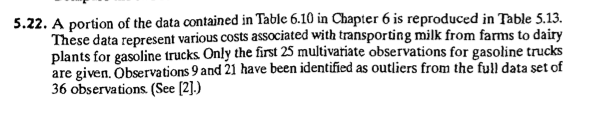
proportions appear close from the point estimate, the difference between them could range from

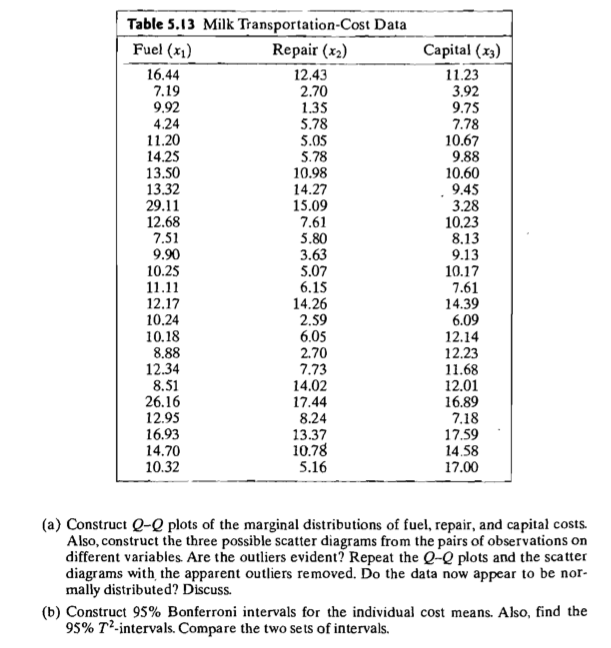
.16 (in favor of the competitor) to about-.1 (in favor of Shorewood). Because the confidence

interval is so wide, I would suggest a larger sample if this information is important to

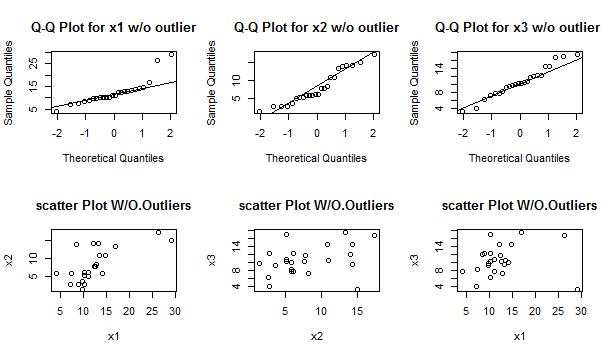
Shorewood.

**Problem 5-22:**

****

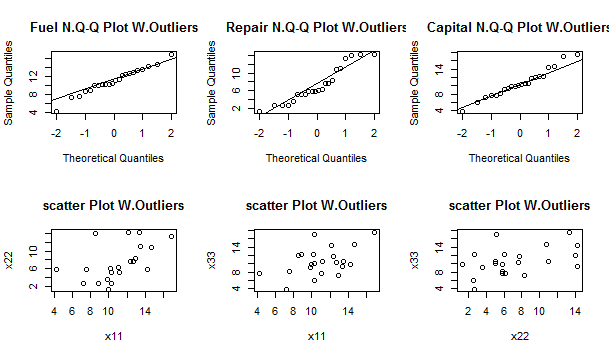
****

Q-Q plots:



From the Q-Q plots and scatterplots, we can find that in x1, the outliers are evident, in x3 the outliers are a little evident, in x2, outliers are not evident. Remove (16.44 12.43 11.23), (29.11 15.09 3.28),(26.16 17.44 16.89).

After removing, Q-Q plots and scatter plots:



The data now appear to be normally distributed. The data distribute along the line.

(b)

The Bonferroni confidence interval for the set consisting of four mean lengths has the following equations:



The Bonferroni confidence interval is

[1,] 8.915009 13.11135

[2,] 4.235914 10.22500

[3,] 8.089353 13.02065



From text, section 5.4, the T2 confidence intervals have the following formulas:



[1,] 9.067712 12.95865

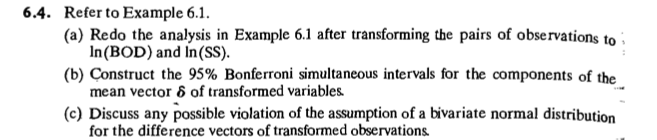
[2,] 4.453854 10.00706

[3,] 8.268801 12.84120



The two set of intervals have little difference, but T2 intervals are smaller than Bonferroni intervals.

**Problem 6-4:**

****

1. The data after transforming,

| **V1** | | **V2** | | **V3** | | **V4** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | |  | |  |
| **1** | 0.7781513 | | 1.431364 | | 1.397940 | | 1.176091 |
| **2** | 0.7781513 | | 1.361728 | | 1.447158 | | 1.113943 |
| **3** | 1.2552725 | | 1.806180 | | 1.556303 | | 1.342423 |
| **4** | 0.9030900 | | 1.643453 | | 1.544068 | | 1.462398 |
| **5** | 1.0413927 | | 1.477121 | | 1.176091 | | 1.491362 |
| **6** | 1.5314789 | | 1.875061 | | 1.643453 | | 1.806180 |
| **7** | 1.4471580 | | 1.414973 | | 1.623249 | | 1.477121 |
| **8** | 1.8512583 | | 2.093422 | | 1.732394 | | 1.806180 |
| **9** | 1.6334685 | | 1.732394 | | 1.531479 | | 1.748188 |
| **10** | 1.5185139 | | 1.477121 | | 1.462398 | | 1.301030 |
| **11** | 1.3010300 | | 1.146128 | | 1.591065 | | 1.322219 |

The T2-statistic for testing H0: is constructed from the differences of paired observations:

| **V1** | | **V2** | |
| --- | --- | --- | --- |
|  |  | |  |
| **1** | **-0.61978876** | | **0.25527251** |
| **2** | **-0.66900678** | | **0.24778448** |
| **3** | **-0.30103000** | | **0.46375729** |
| **4** | **-0.64097806** | | **0.18105468** |
| **5** | **-0.13469857** | | **-0.01424044** |
| **6** | **-0.11197376** | | **0.06888129** |
| **7** | **-0.17609126** | | **-0.06214791** |
| **8** | **0.11886459** | | **0.28724171** |
| **9** | **0.10198954** | | **-0.01579427** |
| **10** | **0.05611594** | | **0.17609126** |
| **11** | **-0.29003461** | | **-0.17609126** |

Taking , we find that , since T2=10.21>9.45, so we still reject H0

(b)

The Bonferroni confidence interval for the set consisting of four mean lengths has the following equations:



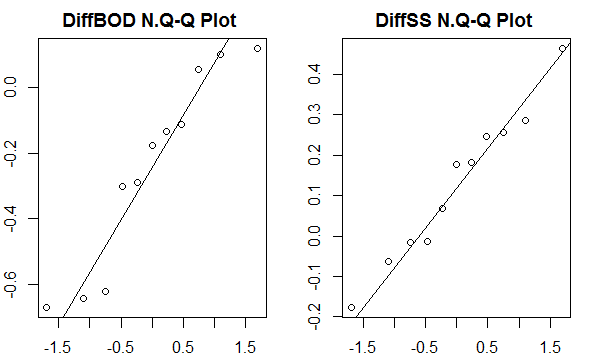
[1] -0.475330077 -0.009512055

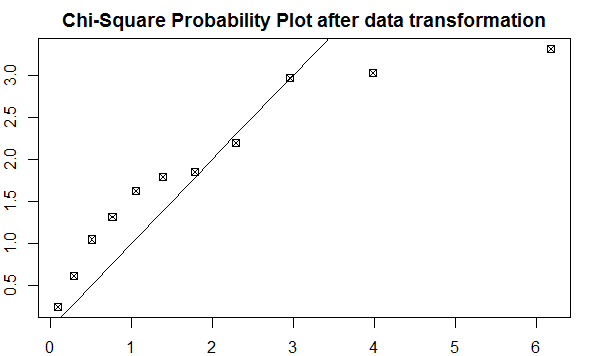
[1] -0.01953654 0.27622915



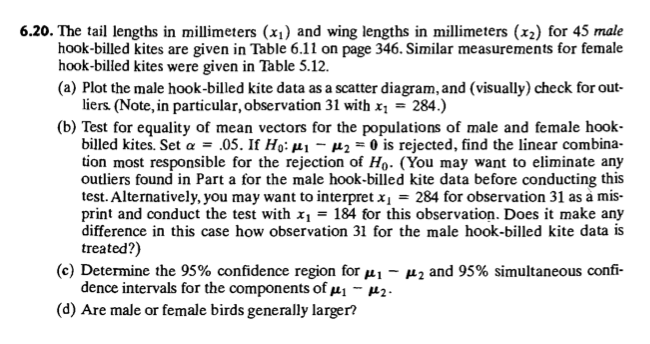
(C)

The Q-Q plots for differences of BOD and difference of SS are shown below. It looks like that both are normally distributed. But Chi-Square Probability Plot is not straight. Although the sample size (n =11) is small, it is difficult to argue for bivariate normality.



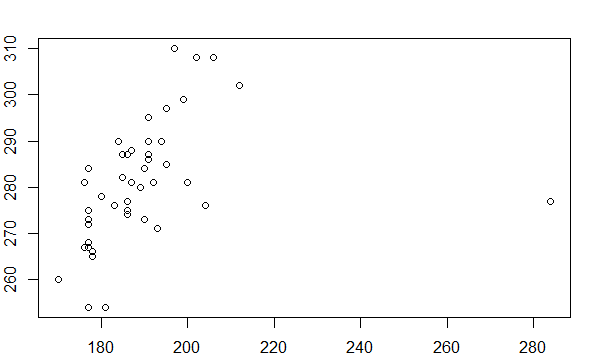


**Problem 6-20:**

****

(a)

Q-Q plots



There is one obvious outliers (24,277),

(b)



So, do not reject null hypothesis, nonzero mean difference does not exist

After removing the outliers, mean vector of population one [187.1591 280.9545], mean vector of population two [193.6222 279.7778]



So, reject equality of mean vectors

The coeffcient of the linear combination of most responsible for rejection is

-0.1566141 0.09342743

Before removing the outliers,

T Squared Based Simultaneous CI for difference

Estimate LowerCI UpperCI

1 -4.311111 -13.016138 4.393916

2 1.088889 -6.925802 9.103580

Bonferroni Based Simultaneous CI for difference

Estimate LowerCI UpperCI

1 -4.311111 -12.187527 3.565305

2 1.088889 -6.162902 8.340680

So, the T2 confidence interval is [-13.016138, 4.393916] and [-6.925802 9.10358]

Boferroni confidence interval is [-12.187527, 3.565305] and [-6.162902,8.34068]

After removing the outliers

T Squared Based Simultaneous CI for the difference

Estimate LowerCI UpperCI

1 -6.463131 -11.898863 -1.027400

2 1.176768 -6.162324 8.515859

Bonferroni Based Simultaneous CI for the difference

Estimate LowerCI UpperCI

1 -6.463131 -11.411601 -1.514661

2 1.176768 -5.504445 7.857980

So, the T2 confidence interval is [-11.898863, -1.0274] and [-6.162324 8.515859]

Boferroni confidence interval is [-11.411601, -1.514661] and [-5.504445,7.857980]

(d)

Female birds are generally larger, since the confidence intervals after removing the outiers for difference in Tails (Male - Female) are negative and the confidence interval for difference in Wings includes zero, indicating no significance difference.

**Problem 6-24:**

Wilks’ lambda 

Where . *S1, S2* and *S3* are sample covariance matrices for the 3 time periods, respectively and *n1, n2* and *n3* are sample sizes ( all equal 30). This matrix measures the residual error. 

Where  is the overall mean (includes all 3 time periods) and the are the individual mean vectors for the 3 time periods. This matrix measures the treatment effect.



The test statistic is given in table 6.3 of the book, using the case where p ≥ 1 and g = 3.



The critical F for this test statistic is 1.99 so we reject the hypothesis that all the means are equal.

To construct the simultaneous confidence intervals, I used the method given in the book section 6.5. The confidence interval:



Where *nk* and *nℓ* are the number of observations in the two groups being compared, and *wii* is

the ith diagonal element from the matrix W. (The treatment effects measure the distance of each

treatment mean from the overall mean. The overall mean drops out of the expression above

when we are comparing two treatments.)

The confidence limits for the mean comparisons for different time periods.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Differences time 1 and time 2 | Lower limits | Upper limits |
| X1 | 1 | -4.44231 | 2.442312 |
| X2 | 0.9 | -2.67371 | 4.473706 |
| X3 | 0.1 | -3.68011 | 3.88011 |
| X4 | 0.3 | -2.06142 | 2.661423 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Differences time 1 and time 3 | Lower limits | Upper limits |
| X1 | -3.1 | -6.54231 | 0.342312 |
| X2 | -0.2 | -3.77371 | 3.373706 |
| X3 | 3.1333 | -0.64678 | 6.913443 |
| X4 | -0.03333 | -2.39467 | 2.328089 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Differences time 2 and time 3 | Lower limits | Upper limits |
| X1 | -2.1 | -5.54231 | 1.342312 |
| X2 | -1.1 | -4.67371 | 2.473706 |
| X3 | 3.0333 | -0.64678 | 6.813443 |
| X4 | -0.3333 | -2.39476 | 2.028089 |

As can be seen, all the confidence intervals include zero, so we can't say that any of the differences observed in the variables over time is significant at alpha = .05.

**Final question for this problem: Are the usual MANOVA assumptions realistic for the data?**

Assumptions:

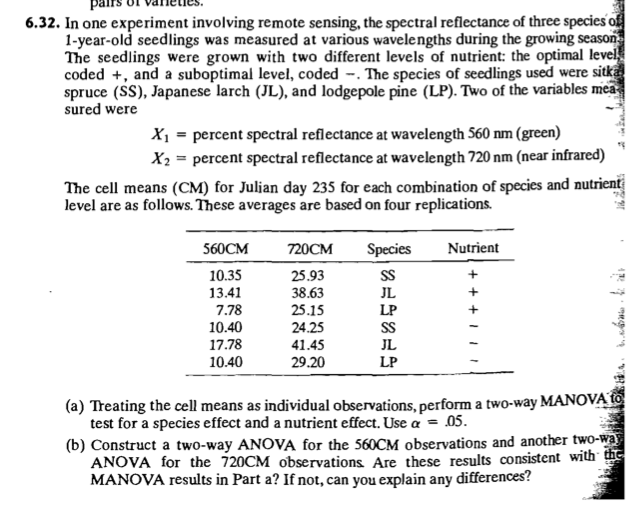
1) The samples are random samples of size *nl* from a p-variate population with mean ug. The random samples from different populations are independent.

2) All populations have a common covariance matrix.

3) Each population is MVN

1. Without knowing more about how the skulls were obtained, it is impossible to say if the samples are random. There could be clusters in the samples (e.g. skulls of a group of family members) or there could be stratification if some researcher over-represented either unusually small or unusually large skulls from a larger collection (although that would probably be unlikely). Or there may have been non-random factors in the historical context that made it more or less likely for some types of people's skulls to be preserved.
2. It is probably reasonable to assume the samples are independent given the large time spreads. Dependence means that the particular skulls included in one group influence the likelihood of skulls in another group being included in that sample. I can't think of a way that would happen, unless the researchers did some very unusual picking and choosing.

**Problem 6-32**

****

(a)

Results from R code:

Df Wilks approx F num Df den Df Pr(>F)

fac1 3 0.53067 3.9758 6 64 0.001923 \*\*

Residuals 33

Df Wilks approx F num Df den Df Pr(>F)

fac2 3 0.26504 10.053 6 64 8.347e-08 \*\*\*

Residuals 33

Two-way MANOVA analysis for both factors separately shows that the P-valueis smaller than 0.05, So we reject the null hypothesis and claims that there is spcies effects and nutrent effects.

(b)

Df Sum Sq Mean Sq F value Pr(>F)

fac1 1 18.1 18.1 0.279 0.600907

fac2 1 1016.7 1016.7 15.669 0.000394 \*\*\*

fac1:fac2 1 1.6 1.6 0.025 0.875716

Residuals 32 2076.4 64.9

Df Sum Sq Mean Sq F value Pr(>F)

fac1 1 3 3 0.019 0.892

fac2 1 4950 4950 34.484 1.57e-06 \*\*\*

fac1:fac2 1 17 17 0.117 0.734

Residuals 32 4594 144

The result shows there is species but no nutrient effect on the 560CM observations, and also species and nutrent both has interacted effect on the 560CM observations.

For the 720CM observations, there is species effect and no nutrient effect on it. However, there is a interacted effect between species and nutrent on the 720CM observations.

**R Code:**

5.10.

```{r}

bear <- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/example1.10.csv")

#Extract the length measurements, find the dimension of the data matrix, and compute the vector of means and the sample covariance matrix:

X=bear[,6:9]

n=dim(X)[1]

p=dim(X)[2]

mean <- sapply(bear[,6:9],mean,na.rm=TRUE)

xbar=matrix(t(mean),ncol=1)

S=cov(X)

#a) Compute the 95% simultaneous T2 confidence intervals for the four population means for length:

c=sqrt((p\*(n-1)/(n-p))\*qf(0.95,p,n-p))

mu.L=xbar-c\*matrix(sqrt(diag(S/n)),ncol=1)

mu.U=xbar+c\*matrix(sqrt(diag(S/n)),ncol=1)

cbind(mu.L,mu.U)

```

(b)

```{r}

#Compute the 95% simultaneous T2 confidence intervals for the increases in mean length from 2 to 3 years

a=matrix(c(-1,1,0,0),ncol=1)

amu.L=t(a)%\*%xbar-c\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+c\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

#Compute the 95% simultaneous T2 confidence intervals for the increases in mean length from 3 to 4 years

a=matrix(c(0,-1,1,0),ncol=1)

amu.L=t(a)%\*%xbar-c\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+c\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

#Compute the 95% simultaneous T2 confidence intervals for the increases in mean length from 3 to 4 years

a=matrix(c(0,0,-1,1),ncol=1)

amu.L=t(a)%\*%xbar-c\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+c\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

```

c)

```{r}

library(ellipse)

a=matrix(c(-1,1,0,0),ncol=1)

b=matrix(c(0,0,-1,1),ncol=1)

meandiff=c(t(a)%\*%xbar,t(b)%\*%xbar)

Sdiff=matrix(c(t(a)%\*%S%\*%a, t(a)%\*%S%\*%b, t(b)%\*%S%\*%a, t(b)%\*%S%\*%b),2,2)

plot(ellipse(Sdiff,centre=meandiff,t=c/sqrt(n)),type="l",xlab="increase 2-3 years",ylab="increase 4-5 years")

points(meandiff[1],meandiff[2])

```

(d)

```{r}

# Construct the 95% Bonferroni confidence intervals for the set consisting of four mean lengths

m=7

cc=qt(0.05/(2\*m),n-1,lower.tail=F)

mu.L=xbar-cc\*matrix(sqrt(diag(S/n)),ncol=1)

mu.U=xbar+cc\*matrix(sqrt(diag(S/n)),ncol=1)

cbind(mu.L,mu.U)

# Construct the 95% Bonferroni confidence intervals for the yearly increas from 2 to three years

a=matrix(c(-1,1,0,0),ncol=1)

amu.L=t(a)%\*%xbar-cc\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+cc\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

# Construct the 95% Bonferroni confidence intervals for the yearly increas from 3 to 4 years

a=matrix(c(0,-1,1,0),ncol=1)

amu.L=t(a)%\*%xbar-cc\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+cc\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

# Construct the 95% Bonferroni confidence intervals for the yearly increas from 4to 5 years

a=matrix(c(0,0,-1,1),ncol=1)

amu.L=t(a)%\*%xbar-cc\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+cc\*sqrt(t(a)%\*%S%\*%a/n)

cbind(amu.L,amu.U)

```

(e)

```{r}

a=matrix(c(-1,1,0,0),ncol=1)

b=matrix(c(0,0,-1,1),ncol=1)

meandiff=c(t(a)%\*%xbar,t(b)%\*%xbar)

Sdiff=matrix(c(t(a)%\*%S%\*%a, t(a)%\*%S%\*%b, t(b)%\*%S%\*%a, t(b)%\*%S%\*%b),2,2)

plot(ellipse(Sdiff,centre=meandiff,t=c/sqrt(n)),type="l",xlab="increase 2-3 years",ylab="increase 4-5 years")

points(meandiff[1],meandiff[2])

a=matrix(c(-1,1,0,0),ncol=1)

amu.L=t(a)%\*%xbar-cc\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+cc\*sqrt(t(a)%\*%S%\*%a/n)

b=matrix(c(0,0,-1,1),ncol=1)

bmu.L=t(b)%\*%xbar-cc\*sqrt(t(b)%\*%S%\*%b/n)

bmu.U=t(b)%\*%xbar+cc\*sqrt(t(b)%\*%S%\*%b/n)

lines(c(amu.L,amu.U),c(bmu.L,bmu.L),lty=2)

lines(c(amu.L,amu.U),c(bmu.U,bmu.U),lty=2)

lines(c(amu.L,amu.L),c(bmu.L,bmu.U),lty=2)

lines(c(amu.U,amu.U),c(bmu.L,bmu.U),lty=2)

c=sqrt((p\*(n-1)/(n-p))\*qf(0.95,p,n-p))

a=matrix(c(-1,1,0,0),ncol=1)

amu.L=t(a)%\*%xbar-c\*sqrt(t(a)%\*%S%\*%a/n)

amu.U=t(a)%\*%xbar+c\*sqrt(t(a)%\*%S%\*%a/n)

b=matrix(c(0,0,-1,1),ncol=1)

bmu.L=t(b)%\*%xbar-c\*sqrt(t(b)%\*%S%\*%b/n)

bmu.U=t(b)%\*%xbar+c\*sqrt(t(b)%\*%S%\*%b/n)

lines(c(amu.L,amu.U),c(bmu.L,bmu.L),lty=2)

lines(c(amu.L,amu.U),c(bmu.U,bmu.U),lty=2)

lines(c(amu.L,amu.L),c(bmu.L,bmu.U),lty=2)

lines(c(amu.U,amu.U),c(bmu.L,bmu.U),lty=2)

```

5.11

(a)

```{r}

Cr=c(0.48,40.53,2.19,0.55,0.74,0.66,0.93,0.37,0.22)

St=c(12.57,73.68,11.13,20.03,20.29,0.78,4.64,0.43,1.08)

# Define the data matrix, find the dimension of the data matrix, and compute the vector of means and the sample covariance matrix:

X=data.frame(cbind(Cr,St))

n=dim(X)[1]

p=dim(X)[2]

mean <- sapply(X,mean,na.rm=TRUE)

xbar=matrix(t(mean),ncol=1)

S=cov(X)

library(ellipse)

plot(ellipse(S,centre=xbar,t=sqrt(((n-1)\*p/(n\*(n-p)))\*qf(0.90,p,n-p))),type="l",xlim=c(-10,25),ylim=c(-10,40))

points(xbar[1,],xbar[2,])

```

(b)

```{r}

mu1.L=xbar[1,]-sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[1,1]/n)

mu1.U=xbar[1,]+sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[1,1]/n)

mu2.L=xbar[2,]-sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[2,2]/n)

mu2.U=xbar[2,]+sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[2,2]/n)

c(mu1.L,mu1.U)

c(mu2.L,mu2.U)

# Plot the confidence intervals together with the confidence ellipse:

lines(c(mu1.L,mu1.L),c(-12,mu2.U),lty=2)

lines(c(mu1.U,mu1.U),c(-12,mu2.U),lty=2)

lines(c(-12,mu1.U),c(mu2.L,mu2.L),lty=2)

lines(c(-12,mu1.U),c(mu2.U,mu2.U),lty=2)

```

(c)

```{r}

#Normal probability plot:

qqnorm(Cr)

qqnorm(St)

# Scatter plot and with ellipse:

plot(X,xlim=c(-20,50),ylim=c(-20,80))

lines(ellipse(S,centre=xbar,level=0.50))

```

(d).

```{r}

XX=X[-2,]

n=dim(XX)[1]

p=dim(XX)[2]

mean <- round(sapply(XX,mean,na.rm=TRUE),2)

xbar=matrix(t(mean),ncol=1)

S=cov(XX)

plot(ellipse(S,centre=xbar,t=sqrt(((n-1)\*p/(n\*(n-p)))\*qf(0.90,p,n-p))),type="l",xlim=c(0,1.5),ylim=c(0,20))

points(xbar[1,],xbar[2,])

mu1.L=xbar[1,]-sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[1,1]/n)

mu1.U=xbar[1,]+sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[1,1]/n)

mu2.L=xbar[2,]-sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[2,2]/n)

mu2.U=xbar[2,]+sqrt(((n-1)\*p/(n-p))\*qf(0.90,p,n-p))\*sqrt(S[2,2]/n)

c(mu1.L,mu1.U)

c(mu2.L,mu2.U)

lines(c(mu1.L,mu1.L),c(-12,mu2.U),lty=2)

lines(c(mu1.U,mu1.U),c(-12,mu2.U),lty=2)

lines(c(-12,mu1.U),c(mu2.L,mu2.L),lty=2)

lines(c(-12,mu1.U),c(mu2.U,mu2.U),lty=2)

```

5.16.

```{r}

p <- c(0.30,0.33,0.16,0.07,0.14)

n <- 355

#a) Compute the 95% simultaneous T2 confidence intervals for the four population means for length:

c=sqrt(qchisq(0.95, 4))

mu.L=p-c\*matrix(sqrt(p\*(1-p)/n),ncol=1)

mu.U=p+c\*matrix(sqrt(p\*(1-p)/n),ncol=1)

cbind(mu.L,mu.U)

# another way

cl <- function(p,n){

CL.S <- p-(sqrt(qchisq(0.95, 4))\*(sqrt(p\*(1-p))/n))

CL.L <- p+(sqrt(qchisq(0.95, 4))\*(sqrt(p\*(1-p))/n))

}

p <- c(0.30,0.33,0.16,0.07,0.14)

n <- 355

for (i in 1:5){

CL.S[i] <- p[i]-(sqrt(qchisq(0.95, 4))\*sqrt(p[i]\*(1-p[i])/355))

CL.L[i] <- p[i]+(sqrt(qchisq(0.95, 4))\*sqrt(p[i]\*(1-p[i])/355))

cl <- cbind(CL.S,CL.L)

}

cl

```

(b)

```{r}

p1 <- 0.30

p2 <- 0.33

t <- sqrt((p1\*(1-p1)+p2\*(1-p2)-2\*p1\*p2)/(355))

cl.s <- (p1-p2)-sqrt(qchisq(0.95,4))\*t

cl.l <- (p1-p2)+sqrt(qchisq(0.95,4))\*t

cbind(cl.s,cl.l)

```

5.22

```{r}

milkT <- read.delim("C:/Users/Administrator/Desktop/guyu/7840/HW3/T5-13.dat", sep="")

x1=milkT[,1]

x2=milkT[,2]

x3=milkT[,3]

qqnorm(x1,main = "Q-Q Plot for x1 w/o outlier")

qqline(x1)

qqnorm(x2,main = "Q-Q Plot for x2 w/o outlier")

qqline(x2)

qqnorm(x3,main = "Q-Q Plot for x3 w/o outlier ")

qqline(x3)

plot(x1,x2,main="scatter Plot W/O.Outliers")

plot(x2,x3,main="scatter Plot W/O.Outliers")

plot(x1,x3,main="scatter Plot W/O.Outliers")

# check the outliers

#install.packages("outliers")

library("outliers")

outlier(milkT)

#remove the outliers

milkT1 <- subset(milkT, x1!=29.11&x2!=17.44&x3!=3.28)

par(mfrow=c(2,3))

x11=milkT1[,1]

x22=milkT1[,2]

x33=milkT1[,3]

qqnorm(x11,main="Fuel N.Q-Q Plot W.Outliers")

qqline(x11)

qqnorm(x22,main="Repair N.Q-Q Plot W.Outliers")

qqline(x22)

qqnorm(x33, main="Capital N.Q-Q Plot W.Outliers")

qqline(x33)

plot(x11,x22,main="scatter Plot W.Outliers")

plot(x11,x33,main="scatter Plot W.Outliers")

plot(x22,x33,main="scatter Plot W.Outliers")

```

(b)

```{r}

#Construct 95% Bonferroni intervals for the individual cost means

milkT2 <- data.frame(milkT1)

m=dim(milkT2)[1]

X=milkT2[,1:3]

n=dim(X)[1]

p=dim(X)[2]

mean <- sapply(milkT2[,1:3],mean,na.rm=TRUE)

xbar=matrix(t(mean),ncol=1)

S=cov(X)

cc=qt(0.05/(2\*m),n-1,lower.tail=F)

mu.L=xbar-cc\*matrix(sqrt(diag(S/n)),ncol=1)

mu.U=xbar+cc\*matrix(sqrt(diag(S/n)),ncol=1)

cbind(mu.L,mu.U)

#the 95% T2-intervals Simultaneous confidence

c=sqrt((p\*(n-1)/(n-p))\*qf(0.95,p,n-p))

mu.L=xbar-c\*matrix(sqrt(diag(S/n)),ncol=1)

mu.U=xbar+c\*matrix(sqrt(diag(S/n)),ncol=1)

cbind(mu.L,mu.U)

```

(c)

```{r}

install.packages("IndependenceTests")

library(IndependenceTests)

X <- table(milkT2)

A.dep.tests(X)

```

From the P-value, which all of them larger than 0.05, none of them are independent.

6.4.

```{r}

#read data in and transform them to log10

wastewater <- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T6-1.txt",header=FALSE, sep="")

wastewater1=log10(wastewater)

# Create dataframe with the differences:

D=as.matrix(cbind(wastewater1[,1]-wastewater1[,3],wastewater1[,2]-wastewater1[,4]))

names(D)=c("bod","ss")

# Compute Hotelling's T2 statistic for testing the null hypothesis d = 0

n=dim(D)[1]

p=dim(D)[2]

dbar=as.matrix(cbind(mean(D[,1]),mean(D[,2])))

Sd=cov(D)

Sdinv <- solve(Sd)

T2=n\*dbar%\*%Sdinv%\*%t(dbar)

T2

# Obtain upper 5% percentile for Hotelling's T2

f.t2=(p\*(n-1)/(n-p))\*qf(0.05,p,n-p,lower.tail=F)

f.t2

# Compute P-value

pval=pf(T2\*(n-p)/((n-1)\*p),p,n-p,lower.tail=F)

pval

```

(b)

```{r}

# Plot the 95% confidence ellipse for the vector of mean differences

library(ellipse)

plot(ellipse(Sd,centre=dbar,t=sqrt(((n-1)\*p/(n\*(n-p)))\*qf(0.95,p,n-p))),type="l")

points(dbar[,1],dbar[,2])

wastewater1=log10(wastewater)

D=as.matrix(cbind(wastewater1[,1]-wastewater1[,3],wastewater1[,2]-wastewater1[,4]))

names(D)=c("bod","ss")

n=dim(D)[1]

p=dim(D)[2]

dbar=as.matrix(cbind(mean(D[,1]),mean(D[,2])))

S=cov(X)

# Compute the 95% Bonferroni confidence intervals for the mean differences:

delta1.LB=dbar[,1]-qt(0.05/(2\*p),n-1,lower.tail=F)\*sqrt(Sd[1,1]/n)

delta1.UB=dbar[,1]+qt(0.05/(2\*p),n-1,lower.tail=F)\*sqrt(Sd[1,1]/n)

delta2.LB=dbar[,2]-qt(0.05/(2\*p),n-1,lower.tail=F)\*sqrt(Sd[2,2]/n)

delta2.UB=dbar[,2]+qt(0.05/(2\*p),n-1,lower.tail=F)\*sqrt(Sd[2,2]/n)

c(delta1.LB,delta1.UB)

c(delta2.LB,delta2.UB)

# Plot the Bonferroni confidence intervals together with the confidence ellipse:

# Compute 95% simultaneous confidence intervals for the two mean differences:

delta1.L=dbar[,1]-sqrt(((n-1)\*p/(n-p))\*qf(0.95,p,n-p))\*sqrt(Sd[1,1]/n)

delta1.U=dbar[,1]+sqrt(((n-1)\*p/(n-p))\*qf(0.95,p,n-p))\*sqrt(Sd[1,1]/n)

delta2.L=dbar[,2]-sqrt(((n-1)\*p/(n-p))\*qf(0.95,p,n-p))\*sqrt(Sd[2,2]/n)

delta2.U=dbar[,2]+sqrt(((n-1)\*p/(n-p))\*qf(0.95,p,n-p))\*sqrt(Sd[2,2]/n)

c(delta1.L,delta1.U)

c(delta2.L,delta2.U)

# Plot the confidence intervals together with the confidence ellipse:

```

(c)

```{r}

par(mfrow=c(1,2))

qqnorm(D[,1],main="DiffBOD N.Q-Q Plot ")

qqline(D[,1])

qqnorm(D[,2],main="DiffSS N.Q-Q Plot ")

qqline(D[,2])

# Chis-Q-Q plot

d <- NULL

for (i in 1:nrow(D)){

d<-cbind(d, t(D[i,] - apply(D, 2, mean))

%\*% solve(var(D))%\*%

(D[i,]-apply(D,2,mean)))

}

# Compute quantiles of a chi-square distribution

q1 <- NULL

for (i in 1:nrow(D)){

q1 <- cbind(q1, qchisq((i-0.5) / (nrow(D)), ncol(D)))

}

# Order the squared distances from smallest to largest

d <- sort(d)

# Specify the plot

par(mar = rep(2, 4))

par(mfrow = c(1,1))

#Create the chi-squared probability plot

d <- matrix(d, nrow = nrow(D), ncol = 1)

q1 <- matrix(q1, nrow = nrow(D), ncol = 1)

plot(q1, d, type="p", pch=7, xlab="Chi-square quantiles",

ylab="Ordered distances", main="Chi-Square Probability Plot after data transformation")

lines(q1, q1, lty=1)

```

6.20.

(a)

```{r}

Mkites<- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T6-11.txt",header=FALSE, sep="")

x1=Mkites[,1]

x2=Mkites[,2]

plot(x1,x2)

```

(b)

```{r}

# the same code in question(c)

```

(c)

```{r}

Mkites<- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T6-11.txt",header=FALSE, sep="")

x1=Mkites[,1]

x2=Mkites[,2]

Fkites <- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T5-12.txt",header=FALSE, sep="")

y1=Fkites[,1]

y2=Fkites[,2]

#function to compare two groups without removing 31 data

paired<-function (x1, x2, level)

{

p <- ncol(x1)

n <- nrow(x1)

d <- x1 - x2

dbar <- apply(d, 2, mean)

s <- cov(d)

tsq <- n \* t(dbar) %\*% solve(s) %\*% dbar

csq <- (n - 1) \* p/(n - p) \* qf(level, p, n - p)

print(tsq)

print(csq)

if (tsq > csq)

cat("\n reject null hypothesis, nonzero mean difference exists \n")

else cat("do not reject null hypothesis, nonzero mean difference does not exist\n")

scit <- matrix(rep(0, p \* 3), nrow = p)

scib <- matrix(rep(0, p \* 3), nrow = p)

for (i in 1:p) {

scit[i, 1] <- dbar[i]

scit[i, 2] <- dbar[i] - sqrt(s[i, i]/n \* csq)

scit[i, 3] <- dbar[i] + sqrt(s[i, i]/n \* csq)

scib[i, 1] <- dbar[i]

scib[i, 2] <- dbar[i] - qt(1 - (1 - level)/(2 \* p), n -

1) \* sqrt(s[i, i]/n)

scib[i, 3] <- dbar[i] + qt(1 - (1 - level)/(2 \* p), n -1) \* sqrt(s[i, i]/n)

}

scit <- data.frame(Estimate = scit[, 1], LowerCI = scit[,

2], UpperCI = scit[, 3])

scib <- data.frame(Estimate = scib[, 1], LowerCI = scib[,

2], UpperCI = scib[, 3])

cat("\n T Squared Based Simultaneous CI for difference \n")

print(scit)

cat("\n Bonferroni Based Simultaneous CI for difference \n")

print(scib)

}

paired(Mkites,Fkites,0.95)

# comapre both after removing the data

Mkites1 <- subset(Mkites,x1<280)

# function to do comparison

twopop<-function (x1, x2, level)

{

p <- ncol(x1)

n1 <- nrow(x1)

n2 <- nrow(x2)

x1bar <- apply(x1, 2, mean)

x2bar <- apply(x2, 2, mean)

cat("\n mean vector of population one \n", x1bar)

cat("\n\n mean vector of population two \n", x2bar)

s1 <- cov(x1)

s2 <- cov(x2)

s.pool <- (n1 - 1)/(n1 + n2 - 2) \* s1 + (n2 - 1)/(n1 + n2 -2) \* s2

tsq <- t(x1bar - x2bar) %\*% solve((1/n1 + 1/n2) \* s.pool) %\*%(x1bar - x2bar)

csq <- (n1 + n2 - 2) \* p/(n1 + n2 - p - 1) \* qf(level, p,n1 + n2 - p - 1)

print(tsq)

print(csq)

if (tsq > csq) {

cat("\n\n reject equality of mean vectors\n\n")

cat("The coeffcient of the linear combination \n of most responsible for rejection is \n\n",

solve(s.pool) %\*% (x1bar - x2bar))

}

else cat("\n\n do not reject equality of mean vectors\n\n")

scit <- matrix(rep(0, p \* 3), nrow = p)

scib <- matrix(rep(0, p \* 3), nrow = p)

for (i in 1:p) {

scit[i, 1] <- x1bar[i] - x2bar[i]

scit[i, 2] <- x1bar[i] - x2bar[i] - sqrt(csq) \* sqrt((1/n1 +1/n2) \* s.pool[i, i])

scit[i, 3] <- x1bar[i] - x2bar[i] + sqrt(csq) \* sqrt((1/n1 +1/n2) \* s.pool[i, i])

scib[i, 1] <- x1bar[i] - x2bar[i]

scib[i, 2] <- x1bar[i] - x2bar[i] - qt(1 - (1 - level)/(2 \*

p), n1 + n2 - 2) \* sqrt((1/n1 + 1/n2) \* s.pool[i,i])

scib[i, 3] <- x1bar[i] - x2bar[i] + qt(1 - (1 - level)/(2 \*p), n1 + n2 - 2) \* sqrt((1/n1 + 1/n2) \* s.pool[i,i])

}

scit <- data.frame(Estimate = scit[, 1], LowerCI = scit[,2], UpperCI = scit[, 3])

scib <- data.frame(Estimate = scib[, 1], LowerCI = scib[,2], UpperCI = scib[, 3])

cat("\n\n T Squared Based Simultaneous CI for the difference \n")

print(scit)

cat("\n\n Bonferroni Based Simultaneous CI for the difference \n")

print(scib)

}

twopop(Mkites1,Fkites,0.95)

```

6.32

```{r}

data <- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T6-18.txt",header=FALSE, sep="")

fac1=data[,3]

# check the species factor with wilk's method

data1=manova(cbind(data[,1], data[,2])~fac1)

summary(data1, test="W")

# check the nutrient factor

fac2=data[,4]

data2=manova(cbind(data[,1], data[,2])~fac2)

summary(data2, test="Wilks")

```

(b)

```{r}

data <- read.csv("C:/Users/Administrator/Desktop/guyu/7840/HW3/T6-18.txt", sep="")

x1=data[,1]

x2=data[,2]

fac1=data[,3]

fac2=data[,4]

summary(aov(x1~fac1\*fac2))

summary(aov(x2~fac1\*fac2))

```

**SAS Code:**

6.24 (I have no ideal to solve this problem using R, so I choose to use sas)

ods html close; /\* close previous \*/

ods html; /\* open new \*/

data skulls;

infile '\\spirit.auburn.edu\gzq0002\Desktop\Guyu\skulls.txt';

input x1 x2 x3 x4 time;

run;

data skull1; set skulls; if time = 1;

data skull2; set skulls; if time = 2;

data skull3; set skulls; if time =3;

proc corr cov data = skulls out = cov;

var x1 x2 x3 x4;

proc corr cov data = skull1 out = cov1;

var x1 x2 x3 x4;

proc corr cov data = skull2 out = cov2;

var x1 x2 x3 x4;

proc corr cov data = skull3 out = cov3;

var x1 x2 x3 x4;

/\* one way manova 3 groups one treatment\*/

/\*proc glm data= skulls alpha = .05 ;

 class time;

 model x1 x2 x3 x4 = time / ss3;

 manova h=\_all\_;

 lsmeans time/ cldiff adjust=bon pdiff;\*/

proc univariate data =skull2;

var x1 x2 x3 x4;

Symbol1 V=Dot;

Histogram / Normal(Mu=Est Sigma=Est Fill)

 CFill=Yellow ;

Inset Mean Std / Header='Normal Parameters'

 Position=(95,95) RefPoint=TR;

QQPlot / Normal(Mu=Est Sigma=Est L=1)

 PctlMinor PCTLSCALE ;

Inset Mean Std / Header='Normal Parameters'

 Position=(95,5) RefPoint=BR;

Run;

proc iml;

\* individual means and variances;

use cov1;

read all var{x1 x2 x3 x4} where(\_type\_='COV') into S1;

read all var{x1 x2 x3 x4} where(\_type\_='MEAN') into x1bar;

print s1;

use cov2;

read all var{x1 x2 x3 x4} where(\_type\_='COV') into S2;

read all var{x1 x2 x3 x4} where(\_type\_='MEAN') into x2bar;

print s2;

use cov3;

read all var{x1 x2 x3 x4} where(\_type\_='COV') into S3;

read all var{x1 x2 x3 x4} where(\_type\_='MEAN') into x3bar;

print s3; print x3bar;

n1 = 30; n2 = 30; n3 = 30; n =n1 + n2 + n3;

W = (n1-1)#s1 + (n2-1)#S2 + (n3-1)#S3; print w;

40

spool = w/(n-1);

Xbar = (n1#x1bar` + n2#x2bar` + n3\*x3bar`)/(n1 + n2 + n3);

xbarmat = xbart//x1bar//x2bar//x3bar;

print xbarmat;

\*create B;

x1bdiff = x1bar` - xbar; x2bdiff = x2bar` - xbar; x3bdiff = x3bar` - xbar;

ss1 = x1bdiff\*x1bdiff`; ss2 = x2bdiff\*x2bdiff`; ss3 = x3bdiff\*x3bdiff`;

ss1n = n1#ss1; print ss1n; ss2n = n2#ss2; ss3n = ss3#n3;

B = ss1n + ss2n + ss3n;

detw = det(W); bpw = B + w; detbpw = det(bpw); wilk = detw/detbpw;

print wilk; print W; print B; print bpw;

n = n1 + n2 + n3; p = nrow(xbar); g = 3;

print n; print p; print g;

\*exact test;

srw = sqrt(wilk);

t1 = (n - p - 2)/p; t2 = (1 - srw)/srw; ftest = t1\*t2; print ftest;

df1= 2\*p; df2 = 2\*(n - p - 2);

FCRIT = finv(.95,df1,df2); print fcrit;

\*\*\*\*\* Construct confidence intervals;

\*\*\*\*\* need one for each difference;

nvec = n1||n2||n3;

\*compare 2 to 3; \* run this code for every pairing change n below;

Do i =1 to p;

c1 = (1/nvec[,2]) + (1/nvec[,3]);

c2= 1/(n-g); cm = c1\*c2; wn=W[i,i];

squared = cm\*wn;

plusmin = sqrt(squared);

plusminv = plusminv//plusmin;

end;

print plusminv;

m = p\*g\*(g-1)/2; print m; alpha = .05/(2\*m); print alpha;

\* get mean differences;

diff12 = x1bar - x2bar; diff13 = x1bar - x3bar; diff23= x2bar-x3bar;

print diff12; print diff13; print diff23;

diffmat = diff12//diff13//diff23; diff = diffmat`; print diff;

al = 1-alpha; free = n-g;

tcrit = tinv(al,free); print tcrit;

plusminf=tcrit#plusminv; print plusminf;

ci12l = diff[,1] - plusminf; ci12u = diff[,1]+plusminf;

ci12 = ci12l||ci12u; print ci12;

ci13l = diff[,2] - plusminf; ci13u = diff[,2]+plusminf;

ci13 = ci13l||ci13u; print ci13;

ci23l = diff[,3] - plusminf; ci23u = diff[,3]+plusminf;

ci23 = ci23l||ci23u; print ci23;

\* box tests for covariance matrices;

xx = (n1-1) + (n2-1) + (n3-1);

u1 = 1/(n1-1) + 1/(n2-1) + 1/(n3-1); u2 = 1/xx; print u1 u2;

u3 = u1 -u2; print u3;

u4 = (2\*p\*p + 3\*p -1)/(6\*(p+1)\*(g-1)); print u4;

u = u3\*u4;

bb=log(det(spool)); dd1 = log(det(s1)); dd2 = log(det(s2)); dd3 =

log(det(s3));

v = p\*(p+1)\*(g-1)/2; print v;

Critc = cinv(.95,20); print critc;

d1 = det(s1); d2 = det(s2); d3 = det(s3);

lnd1 = log(d1); lnd2 = log(d2); lnd3 = log(d3);

dp = det(spool); lndp = log(dp); print d1 d2 d3 dp; print u;

41

c1= xx\*lndp - (n1-1)\*lnd1 - (n2-1)\*lnd2 - (n3-1)\*lnd3; print lndp lnd1 lnd2

lnd3;

c= (1-u)\*c1; print c;

/\*\*\*\* mahalanobis distance \*\*\*\*\*\*\*\*/

use cov;

read all var{x1 x2 x3 x4} where(\_type\_='COV') into S;

read all var{x1 x2 x3 x4} where(\_type\_='MEAN') into xbar;

use skulls;

read all var{x1 x2 x3 x4} into x;

sinv = inv(s1); xbart=x1bar`;

do i = 1 to n1;

meanmat = meanmat||xbart;

end;

mmt = meanmat`;

dev = X- mmt; print x; print mmt;

mahal=J(n,1,0);

print dev;

do i = 1 to n;

devi = dev[i,]; devit = devi`;

mahala = devi\*sinv\*devit;

mahal[i,1] = mahala;

end;

call sort(mahal,1);

print mahal;

create mahalad from mahal[colname = {"dsquare"}];

append from mahal;

close mahald;

run;

data new;

do i = 1 to 30;

prob = (i-.5)/30;

output;

end;

data newer; set new;

chis = cinv(prob,4,0);

run;

proc sort data = mahalad; by dsquare;

data plot;

merge newer mahalad;

run;

proc gplot data = plot;

plot chis\*dsquare;

run; quit;