# STAT 530 - Fall 2003 - Sample Homework 3 Solutions

1) The two sets of mean vectors and the three covariance matrices could be entered as follows:

```
n1<-10
n2<-10

xbar1.m1<-c(5,8)
xbar2.m1<-c(6,9)

xbar1.m2<-c(5,8)
xbar2.m2<-c(4,9)

C1<-matrix(c(1,0.5,0.5,1),byrow=T,ncol=2)
C2<-matrix(c(1,0,0,1),byrow=T,ncol=2)
C3<-matrix(c(1,-0.5,-0.5,1),byrow=T,ncol=2)
```

And then we could use the code for calculating Tsquare:

```
Cinv <- solve(C1)
Tsquare <- (n1*n2)*t(xbar1.m1-xbar2.m1)%*%Cinv%*%(xbar1.m1-xbar2.m1)/(n1+n2)
Tsquare

Cinv <- solve(C2)
Tsquare <- (n1*n2)*t(xbar1.m1-xbar2.m1)%*%Cinv%*%(xbar1.m1-xbar2.m1)/(n1+n2)
Tsquare

Cinv <- solve(C3)
Tsquare <- (n1*n2)*t(xbar1.m1-xbar2.m1)%*%Cinv%*%(xbar1.m1-xbar2.m1)/(n1+n2)
Tsquare

Cinv <- solve(C1)
Tsquare <- (n1*n2)*t(xbar1.m2-xbar2.m2)%*%Cinv%*%(xbar1.m2-xbar2.m2)/(n1+n2)
Tsquare

Cinv <- solve(C2)
Tsquare <- (n1*n2)*t(xbar1.m2-xbar2.m2)%*%Cinv%*%(xbar1.m2-xbar2.m2)/(n1+n2)
Tsquare

Cinv <- solve(C3)
Tsquare <- (n1*n2)*t(xbar1.m2-xbar2.m2)%*%Cinv%*%(xbar1.m2-xbar2.m2)/(n1+n2)
Tsquare
```

This gives us the values shown below

| Mean \ Covariance | C1 = positive association | C2=independence | C3=negative association |
|---|---|---|---|
| M1 = uniform case | $T^2$=6.667 | $T^2$=10 | $T^2$=20 |
| M2 = split case | $T^2$=20 | $T^2$=10 | $T^2$=6.667 |

The $T^2$ has the most power when the difference in the means goes against what would be expected from the covariances. That is, when the two variables have a positive correlation it has additional power in the split case, but when the two variables have a negative correlation it has additional power in the uniform case.

2) *This problem continues problem 3 on homework 2, and again uses the data set*
*Use Hotelling's $T^2$ to test whether there is a difference between the male and female bears on variables 3-7.*

After entering the function tsquare,

```
bears<-read.table("http://www.stat.sc.edu/~habing/courses/data/bears.txt",head=T)
bm<-bears[bears[,2]==1,3:7]
bf<-bears[bears[,2]==2,3:7]
tsquare(bm,bf)
```

a) *Report your conclusion at an overall $\alpha$ level of 0.05.*

```
$p.value
           [,1]
  [1,] 0.228367
```

We fail to reject the null hypothesis at the a=0.05 level. We do not have sufficient evidence to say that the male and

b) *Do the assumptions for performing this test seem to be met? Why or why not?*

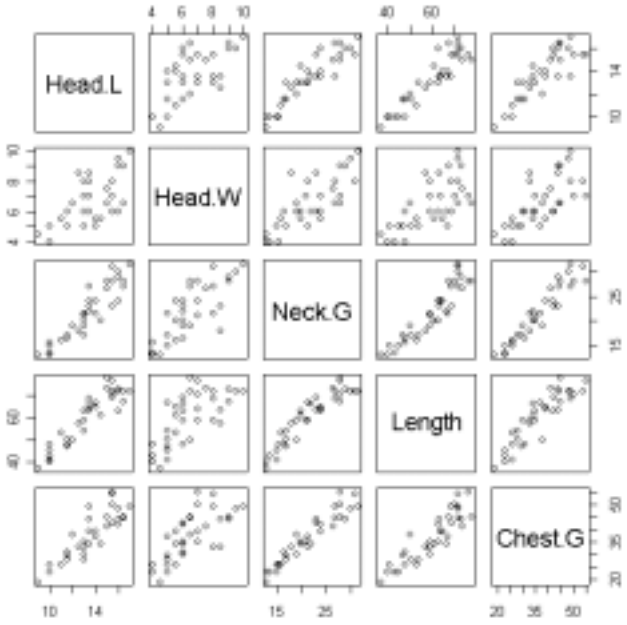
The assumptions of independence and random samples are the same as in hmwk 2. We need to check that the covariance matrices are equal.

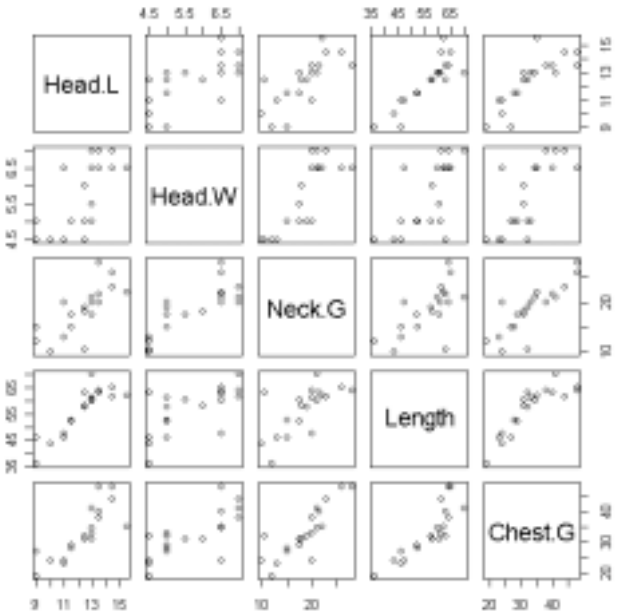
```
> cov(bm)
           Head.L    Head.W    Neck.G    Length  Chest.G
Head.L   4.983135  2.536310 11.415675  23.52052 18.84909
Head.W   2.536310  2.605357  7.045833  12.49845 11.41988
Neck.G  11.415675  7.045833 30.549802  58.79258 50.23544
Length  23.520516 12.498452 58.792579 130.03933 97.67361
Chest.G 18.849087 11.419881 50.235437  97.67361 91.10790

> cov(bf)
           Head.L    Head.W    Neck.G    Length   Chest.G
Head.L   3.049342 1.1480263  6.210526 13.519737 11.523026
Head.W   1.148026 0.9072368  3.763158  5.569737  5.692763
Neck.G   6.210526 3.7631579 23.315789 28.618421 32.934211
Length  13.519737 5.5697368 28.618421 78.160526 61.817105
Chest.G 11.523026 5.6927632 32.934211 61.817105 66.323026
```

Just by visually comparing the values, the covariances seem to be off about as far as the variances are... and so I would probably come to the same conclusion with them that was reached with the variances. So, they seem border-line. Hotellings $T^2$ is fairly robust when the sample sizes are equal, but they aren't here. There are 36 males and 20 females.

The rule of thumb is that if the larger group has the larger generalized variance the procedure will be conservative, and it will be liberal if the opposite is true. Here `det(tsquare(bm,bf)$C1)= 2285.871` and `det(tsquare(bm,bf)$C2) = 1489.693`, so we expect the test to be conservative.

To check multivariate normality we need to check the individual q-q plots (which we did in Homework 2) and also check the pair-wise scatter plots to see if they look ellipsoid.




The only ones that look very questionable are those with the female head width (on the right). But since the sample size is small it is hard to tell. As the procedure is fairly robust and there are no major outliers or overly unusual patterns I would probably not object to the assumption of multivariate normality.

c) *Briefly compare the p-value you observed in part a of this problem to the p-values you for the individual t-tests found in homework 2. In light of what you found in question 1 is this result surprising?*

The p-value for Hotelling's $T^2$ of .228367 s near the large end of those found from the individual t-tests:

```
Head Length Head Width  Neck Girth   Length       Chest Girth
0.09861589 0.06895838 0.03265619 0.22991315 0.15448976
```

Thecorrelation between all of the variables is positive (see the matrices shown in part b). And comparing the means

```
> apply(bm,2,mean)
   Head.L    Head.W    Neck.G    Length   Chest.G
13.347222  6.458333 21.736111 60.230556 36.730556
> apply(bf,2,mean)
 Head.L  Head.W  Neck.G  Length Chest.G
 12.375   5.725  18.500  56.650  33.075
```

we see the males have larger mean on each variable. This is the uniform difference/positive correlation case and so we would expect it to have lower power (smaller $T/T^2$ values and therefore larger p-values).

3a) *Calculate Mahalinobis distance between the sample means of the three species. Which two species are most similar?*

After entering the function `mahsetup`, and noting that the distance matrix will be 3x3 and not 5x5, we could use the following:

```
library(MASS)
iris<-read.table("http://www.stat.sc.edu/~habing/courses/data/iris.txt",head=T)
xbarandC<-mahsetup(iris[,2:5],iris[,6])
xbar<-xbarandC$xbarmatrix
C<-xbarandC$covmatrix
mahmat<-matrix(0,nrow=3,ncol=3)
for (i in 1:3){
  for (j in 1:3){
    mahmat[i,j]<-mahalanobis(xbar[,i],xbar[,j],C)
}}
rownames(mahmat)<-colnames(mahmat)<-colnames(xbar)
mahmat
```

```
              setosa versicolor virginica
setosa       0.00000   89.86419 179.38471
versicolor  89.86419    0.00000  17.20107
virginica  179.38471   17.20107   0.00000
```

We can see that Virginica and Versicolor are the most similar and that both differ from Setosa (especially the Virginica).

b) *Conduct a principal components analysis on this data.*

It is probably best to use the correlation matrix so that we don't have to worry about the scale of the different measurements, or how to weight them.

```
irisdat<-iris[,2:5]
iris.pca<-princomp(irisdat,cor=T)
```

(i) *Report the loadings and give a descriptive name to each of the components.*

```
loadings(iris.pca)
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.521 -0.377  0.720  0.261
Sepal.Width  -0.269 -0.923 -0.244 -0.124
Petal.Length  0.580        -0.142 -0.801
Petal.Width   0.565        -0.634  0.524
```

The first principal component seems to be everything except sepal width weighted about evenly, and contrasted slightly with sepal width. Perhaps "Overall Size (except sepal width)".

The second component seems to be a combination of the two sepal measurements, but especially the sepal width. Because of the negative sign we might call it either "Sepal smallness" or "Lack of Sepal Width". (Or you could multiply through by -1 and give them the slightly easier names "Sepal size" or "Sepal Width".)

The third component is primarily the contrast between Sepal Length and Petal Width, so "Sepal Length vs. Petal Width".

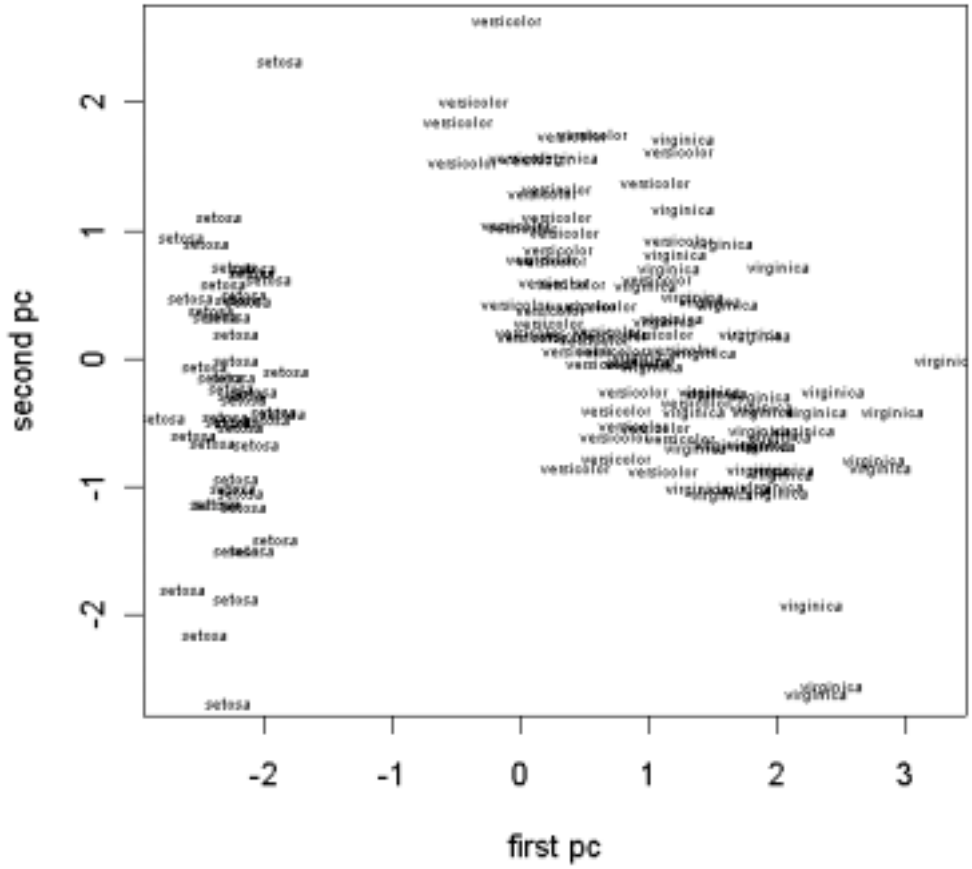
The last one might be "Petal Width vs. Petal Length" or "Petal squatness"?

(ii) *Report the percentage of the variation explained by each of the components*

```
summary(iris.pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4
Standard deviation     1.7083611 0.9560494 0.38308860 0.143926497
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
Cumulative Proportion  0.7296245 0.9581321 0.99482129 1.000000000
```

(iii) *Plot the irises by their first two principal components, using a different symbol for each type of iris. Does the plot seem to confirm the results in part a?*

```
iris.pred<-predict(iris.pca)
eqscplot(iris.pred[,1:2],type="n",xlab="first pc",ylab="second pc")
text(iris.pred[,1:2],labels=as.character(iris[,6]),cex=0.5)
```




This seems to agree with the distances we found. Virginica and Versicolor are close together and both are far from Setosa, with Virginica beting the farthest.

(iv) *How many of the principal components do you think it is worth using? Why?*

From part 2, the proportion of the variance explained rule would lead to keeping two components.

```
Proportion of Variance 0.7296245 0.2285076 0.03668922 0.005178709
```

Looking at the eigen values for the correlation matrix, only one value is larger than one... but the second is fairly close. This would lead to keeping at least one and probably two components.

```
eigen(cor(irisdat))
$values
[1] 2.91849782 0.91403047 0.14675688 0.02071484
```

With only four values, the scree plot is not particularly helpful, but again it seems like there are two factors before it levels out to small values.

```
plot(iris.pca)
```