Department of Statistics
Jahangirnagar University
Professional Masters in Applied Statistics and Data Science (ASDS)
Course Title: Introduction to Data Science with Python
Course No.: PM-ASDS06 (Section: B)
Course Teacher: Dr. Rumana Rois

Assignment 1, Summer 2021

At the start of a study to determine whether exercise or dietary supplements would slow bone loss in older women, an investigator measured the mineral content of bones by photon absorptiometry. Measurements were recorded for three bones on the dominant and nondominant sides and are shown in Table 1. This data is available in the file **multivar5th/T1-8.dat**.

Table 1: Mineral Content in Bones

Subject number	**Dominant radius** (x_1)	**Radius** (x_2)	**Dominant humerus** (x_3)	**Humerus** (x_4)	**Dominant ulna** (x_5)	**Ulna** (x_6)
1	1.103	1.052	2.139	2.238	0.873	0.872
2	0.842	0.859	1.873	1.741	0.590	0.744
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	0.915	0.936	1.971	1.869	0.869	0.868
26	0.89**XX**	0.8**XX**	1.79**XX**	1.7**XX**	0.7**XX**	0.6**XX**
27	0.91**XX**	0.9**XX**	1.81**XX**	1.8**XX**	0.6**XX**	0.8**XX**
28	0.84**XX**	0.9**XX**	1.89**XX**	1.7**XX**	0.7**XX**	0.7**XX**
29	0.88**XX**	0.8**XX**	1.91**XX**	1.8**XX**	0.8**XX**	0.6**XX**
30	0.83**XX**	0.9**XX**	1.94**XX**	1.9**XX**	0.6**XX**	0.8**XX**

XX is the last two digits of your exam roll number.

1. Examine the multivariate normality of the observations on six different variables of the mineral content of three bones on the dominant and nondominant sides of older women. Also, detect the outliers (if any).

```
######## Chi-square Plot ######
X1=matrix(c(126974, 96933, 86656, 63438, 55264, 50976, 39069, 36156, 35209, 32416),nrow=10,ncol = 1)
X2 = matrix(c(4224, 3835, 3510, 3758, 3939, 1809, 2946, 359, 2480, 2413),nrow=10, ncol=1, byrow = TRUE)
X3 = matrix(c(173297, 160893, 83219, 77734, 128344, 39080, 38528, 51038, 34715, 25636), nrow=10, ncol=1)
p=3
X=cbind(X1, X2, X3)
X_bar=colMeans(X)
S = cov(X)
n=length(X1)
Dsq=matrix(0,nrow=n,ncol=1)
for (i in 1:n) {
 Dsq[i] = t(X[i,]-X_bar)%*%solve(S)%*%(X[i,]-X_bar)
}
dj=sort(Dsq, decreasing = F)
j=matrix(seq(1,n), nrow=n, ncol=1)
qj = qchisq((j-0.5)/n, p)  ## USE THIS FOR GETTING THE CORRECTED PLOT
plot(qj,dj)
```

2. Evaluate T^2 of the six variables $(x_1, x_2, \ldots, x_6)$ for testing $H_0 : \mathbf{\mu}' = [0.80 \quad 0.80 \quad 1.70 \quad 1.70 \quad 0.70 \quad 0.70]$ at 5% level of significance. Hence,

find out the sampling distribution of T^2. [Hint: sampling distribution of T^2 is $\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$]

3. Construct the sample covariance matrix S for the above data matrix. Hence, determine the sample **principal components** and their variances for the covariance matrix S. How many principal components will be retained in this analysis?

```
#######Principal Components #############
x=read.table('T6-9.dat',header = FALSE)
turtles=log(x[25:48,1:3]) ### turtles is the data set
xbar=colMeans(turtles)
S=cov(turtles)
fit <- prcomp(S)  #princomp(S)
fit    ### To get the PC after rotation
summary(fit)  ##  To get the proportions
screeplot(fit, npcs = 3, type = "lines")
eigen(S) ### To get the PC without rotation
```

4. Conduct the **factor analysis** with these 6 variables and $m = 2$ common factors using maximum likelihood procedure and find the followings:

(i) Find the estimated factor loadings and communalities.

(ii) What proportion of the total population variance is explained by the first common factors? And by the 2nd common factor.

(iii) Check whether the 2 factors are adequate for our model?

```
fac <- factanal(x, factors=2, method='mle', scale=T, center=T)
fac
factanal(x, factors=2, method='PCA', scale=T, center=T)
```

```
Call:
factanal(x = x, factors = 2, method = "PCA", scale = T, center = T)

Uniquenesses:
   V1    V2    V3    V4    V5
0.497 0.252 0.474 0.610 0.176

Loadings:
   Factor1 Factor2
V1 0.601   0.378
V2 0.849   0.165
V3 0.643   0.336
V4 0.365   0.507
V5 0.207   0.884

               Factor1 Factor2
SS loadings      1.671   1.321
Proportion Var   0.334   0.264
Cumulative Var   0.334   0.598

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.58 on 1 degree of freedom.
The p-value is 0.448
```

5. Calculate the Euclidean distances between six different variables of the mineral content of three bones on the dominant and nondominant sides of older women. **Cluster** the six variables using the single linage and complete linkage hierarchical methods. Draw the dendrograms and compare the results.