



Stat lab 413

NAME:AL ABID MD AHAD ISLAM

EXAM ROLL: 210445

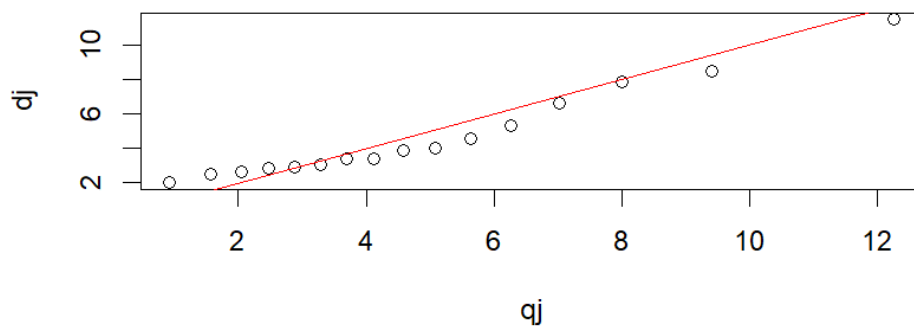
SESSION:2020-2021

1. **Examine the multivariate normality of the observations on five types of overtime hours for the Madison, Wisconsin, Police department.**

Code:

```
a=T5.8
a
x=as.matrix(a)
n=length(x[,1])
S=cov(x)
solve(S)
x_bar=colMeans(x)
x_bar=matrix(colMeans(x),nrow=1,ncol=5)
d_sq=matrix(0,nrow=n,ncol=1)
for(i in 1:n){
  d_sq[i]=(x[i,]-x_bar)%*%solve(S)%*%t(x[i,]-x_bar)
}
d_sq
dj=sort(d_sq,decreasing=FALSE)
dj
j=matrix(seq(1,n),ncol=1,nrow=n)
qj=qchisq((j-0.5)/n,5)
qj
plot(qj,dj)
abline(0,1,col='red')
```

output:



Interpretation:

this is a chi square plot. We see that the data point are far from the red line(45 degree line) which suggests non-normality for the data set.

2. Evaluate T2 of the six variables (x1,x2,...,x5) for testing $H_0: \mu = [1400 \ 2600 \ 13500 \ 800]$. Hence, find out the sampling distribution of T2 .

Code:

```
mu0=c(3500 ,1400, 2600, 13500 ,800)
S=cov(a)
X_bar=colMeans(a)
n=80 ###col*row
p=5
T_2=n*t(X_bar-mu0)%*%solve(S)%*%(X_bar-mu0)

Lcv= ((n-1)*p/(n-p))*qf(.05, p, (n-p))
Ucv= ((n-1)*p/(n-p))*qf(.95, p, (n-p))
## if T2< Lcv or T2 >Ucv2, then H0 will be rejected
##H0=the means are equal
if (T_2<Lcv | T_2>Ucv) print("H0 is rejected") else print("H0 is not rejected")
```

output:

```
S
      V1      V2      V3      V4      V5
V1 367884.73 -72093.82  85714.77 222491.4 -44908.33
V2 -72093.82 1399053.06  43399.86 139692.2 110517.13
V3  85714.77  43399.86 1458543.05 -1113809.8 330923.80
V4 222491.43 139692.24 -1113809.78 1698324.4 -244785.87
V5 -44908.33 110517.13  330923.80 -244785.9  224718.00
```

```
T_2
      [,1]
[1,] 2.190693
Lcv
[1] 1.19202
Ucv
[1] 12.30597
"H0 is not rejected"
```

Interpretation:

As the calculated value is in the acceptance region . so, we will say that we failed to reject the null hypothesis.

3. Construct the principal component analysis using the sample covariance matrix S for the above data matrix.

- Determine the sample principal components and their variances for the covariance matrix S.
- Compute the proportion of total variance explained by the first two principal components. Interpret your result.

Code:

```
fit=princomp(a,cor = FALSE)
summary(fit)
```

output:

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1664.3997	1195.4942	792.5457	470.2534	315.9490
Proportion of Variance	0.5381	0.2776	0.1220	0.0429	0.0193
Cumulative Proportion	0.5381	0.8157	0.9377	0.9806	1.0000

> (fit)

Standard deviations (1, ..., p=5):

[1] 1664.3997 1195.4942 792.5457 470.2534 315.9490

Rotation (n x k) = (5 x 5):

	PC1	PC2	PC3	PC4	PC5
V1	0.04623685	-0.04824957	-0.62915050	-0.6428767	-0.431756103
V2	0.03900887	0.98481068	0.07669642	-0.1506621	0.006694838
V3	-0.65824478	0.10702841	-0.58181284	0.2504276	0.392477505
V4	0.73412440	0.06940658	-0.50323172	0.3970439	0.212974429
V5	-0.15529213	0.10745954	-0.08086384	0.5862190	-0.783674134

Interpretation:

i)Principal Components:

PC1 = 0.0462V1 + 0.0390V2 - 0.6582V3 + 0.7341V4 - 0.1553V5

PC2 = -0.0482V1 + 0.9848V2 + 0.1070V3 + 0.0694V4 + 0.1075V5

PC3 = -0.6292V1 + 0.0767V2 - 0.5818V3 - 0.5032V4 - 0.0809V5

PC4 = -0.6429V1 - 0.1507V2 + 0.2504V3 + 0.3970V4 + 0.5862V5

PC5 = -0.4318V1 + 0.0067V2 + 0.3925V3 + 0.2130V4 - 0.7837V5

Variances of Principal Components:

Var(PC1) = 2770226.36

Var(PC2) = 1429206.38

Var(PC3) = 628128.69

Var(PC4) = 221138.26

Var(PC5) = 99823.77

ii)Proportion of Variance Explained:

PC1 = 0.5381

PC2 = 0.2776

Total (PC1 + PC2) = 0.8157 (81.57%)

The first two principal components explain 81.57% of the total variance. Thus, most of the variability in the data can be captured using only two components, allowing effective dimensionality reduction from 5 variables to 2.

4. Conduct the factor analysis with 5 variables (five types of overtime hours for the Madison, Wisconsin, Police department) and 2=m common factors.

(i). Find the matrix of specific variances. Hence, define the most significant variable which fit neatly into our factors.

(ii). Find the estimated factor loadings and communalities. Interpret the estimated factor loadings.

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

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

Code:

```
fac2 = factanal(a, factors=2,method='mle', scale=T, center=T)
```

fac2

output:

Call:

```
factanal(x = a, factors = 2, method = "mle", scale = T, center = T)
```

Uniquenesses:

	V1	V2	V3	V4	V5
	0.005	0.988	0.005	0.362	0.614

Loadings:

	Factor1	Factor2
v1	0	0.997
v2	0	-0.101
v3	0.992	0.102
v4	-0.743	0.294
v5	0.599	-0.166

	Factor1	Factor2
SS loadings	1.896	1.129
Proportion Var	0.379	0.226
Cumulative var	0.379	0.605

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 1.5 on 1 degree of freedom.
The p-value is 0.221

Interpretation:

Matrix of Specific Variances (Uniquenesses)

Psi =

```
[0.005 0 0 0 0]
```

```
[0  0.988 0  0  0]
[0  0  0.005 0  0]
[0  0  0  0.362 0]
[0  0  0  0  0.614]
```

Factor Loadings Matrix

```
L =
[ 0  0.997 ]
[ 0 -0.101 ]
[ 0.992 0.102 ]
[-0.743 0.294 ]
[ 0.599 -0.166 ]
```

Communalities:

$H_i^2 = \text{sum of squared loadings}$

```
V1 = 0.994
V2 = 0.010
V3 = 0.994
V4 = 0.638
V5 = 0.386
```

Variance Explained

```
Factor 1: 37.9%
Factor 2: 22.6%
Total: 60.5%
```

Model Adequacy

Chi-square = 1.5, df = 1, p-value = 0.221.

The p value is greater than 0.05 ,so the null hypothesis(2 factors are adequate for our model) is failed to reject.

Conclusion: 2-factor model is adequate.

5. Calculate the sample correlation matrix R for the above data matrix. Arrange the data into two sets of variables, i.e.,)1(X = { 1 X , 2 X , 3 X } and)2(X = { 4 X , 5 X }.

(i) Find all the sample canonical correlations and all the pairs of sample canonical variates. Hence, interpret the first sample canonical variates $1^{\wedge}U$ and $1^{\wedge}V$.

(ii) Let) Z 1 and Z (2) be the sets of standardized variables corresponding to X) (X (2) (1 , respectively. What proportion of the total sample variance of the first set is explained by the canonical variate $U^{\wedge} 1$ and Z (1) ? What proportion of the total sample variance of the Z (2) set is explained by the canonical variate $1^{\wedge} V$?

Code:

```
library(CCA)

R <- cor(a)

print("Sample correlation matrix:")

print(round(R, 3))

X1 <- a[, 1:3]

X2 <- a[, 4:5]

cca_result <- cc(X1, X2)

print(paste("Canonical correlations:", paste(round(cca_result$cor, 4), collapse = " , ")))

print("(i) First canonical variates:")

print("U1 coefficients:")

print(round(cca_result$xcoef[,1], 4))

print("V1 coefficients:")

print((cca_result$ycoef[,1]))


print("(ii) Proportion of variance explained:")

Z1 <- scale(X1); Z2 <- scale(X2)

U1 <- as.matrix(Z1) %*% cca_result$xcoef[,1]

V1 <- as.matrix(Z2) %*% cca_result$ycoef[,1]

prop_U1 <- sum(cor(Z1, U1)^2) / ncol(Z1)

prop_V1 <- sum(cor(Z2, V1)^2) / ncol(Z2)

print(paste("Variance in Z(1) explained by U1:", round(prop_U1, 4)))

print(paste("Variance in Z(2) explained by V1:", round(prop_V1, 4)))
```

output:

```
library(CCA)
> R <- cor(a)
> print("Sample correlation matrix:")
[1] "Sample correlation matrix:"
> print(round(R, 3))
      v1      v2      v3      v4      v5
v1  1.000 -0.100  0.117  0.281 -0.156
v2 -0.100  1.000  0.030  0.091  0.197
v3  0.117  0.030  1.000 -0.708  0.578
v4  0.281  0.091 -0.708  1.000 -0.396
v5 -0.156  0.197  0.578 -0.396  1.000
> x1 <- a[, 1:3]
> x2 <- a[, 4:5]
> cca_result <- cc(x1, x2)
> print(paste("Canonical correlations:", paste(round(cca_result$cor, 4), collapse = ", ")))
[1] "Canonical correlations: 0.8639, 0.2836"
> print("(i) First canonical variates:")
[1] "(i) First canonical variates:"
> print("U1 coefficients:")
[1] "U1 coefficients:"
> print(round(cca_result$xcoef[,1], 4))
      v1      v2      v3
-7e-04 -1e-04  8e-04
> print("V1 coefficients:")
[1] "V1 coefficients:"
> print((cca_result$ycoef[,1]))
      v4      v5
-0.0005943896  0.0008354843
>
> print("(ii) Proportion of variance explained:")
[1] "(ii) Proportion of variance explained:"
> Z1 <- scale(x1); Z2 <- scale(x2)
> U1 <- as.matrix(Z1) %*% cca_result$xcoef[,1]
> V1 <- as.matrix(Z2) %*% cca_result$ycoef[,1]
> prop_U1 <- sum(cor(Z1, U1)^2) / ncol(Z1)
> prop_V1 <- sum(cor(Z2, V1)^2) / ncol(Z2)
> print(paste("Variance in Z(1) explained by U1:", round(prop_U1, 4)))
[1] "Variance in Z(1) explained by U1: 0.2943"
> print(paste("Variance in Z(2) explained by V1:", round(prop_V1, 4)))
[1] "Variance in Z(2) explained by V1: 0.6933"
```

Interpretation:

Sample Correlation Matrix (R)

R =

```
[1.000 -0.100  0.117  0.281 -0.156]
[-0.100  1.000  0.030  0.091  0.197]
[0.117  0.030  1.000 -0.708  0.578]
[0.281  0.091 -0.708  1.000 -0.396]
[-0.156  0.197  0.578 -0.396  1.000]
```


Canonical Correlations

$$\rho_1 = 0.8639$$

$$\rho_2 = 0.2836$$

First Canonical Variates

$$U_1 = -0.0007V_1 - 0.0001V_2 + 0.0008V_3$$

$$V_1 = -0.000594V_4 + 0.000835V_5$$

The first canonical variates show that V3 (from first set) and V5 (from second set) contribute the most. The canonical correlation (0.8639) indicates a strong relationship between the two sets of variables.

Proportion of Variance Explained

Variance in first set explained by U1: 0.2943 (29.43%)

Variance in second set explained by V1: 0.6933 (69.33%)

The first canonical correlation is strong, but it explains more variance in the second set than in the first set. The second canonical correlation is relatively weak and less important.

6. Calculate the Euclidean distances between five different variables of overtime hours. Cluster the five variables using the single linkage and complete linkage hierarchical methods. Draw the dendrograms and compare the results.

Code:

```
Z <- scale(a)
```

```
dist.vars=dist(t(Z), method = "euclidean")
```

```
hc_single=hclust(dist.vars, method = "single")
```

```
hc_complete <- hclust(dist.vars, method = "complete")
```

```
# Plot dendrograms
```

```
par(mfrow=c(1,2))
```

```
plot(hc_single, main = "Single linkage clustering of variables", labels =
```

```
  colnames(a))
```

```
plot(hc_complete, main = "Complete linkage clustering of variables", labels =
```

```
  colnames(a))
```

```
par(mfrow=c(1,1))
```

```
clusters_single <- cutree(hc_single, k = 2)
```

```
clusters_complete <- cutree(hc_complete, k = 2)
```

```
clusters_single; clusters_complete
```

output:

```
> Z <- scale(a)
> dist.vars=dist(t(Z), method = "euclidean")
> hc_single=hclust(dist.vars, method = "single")
> hc_complete <- hclust(dist.vars, method = "complete")
>
> # Plot dendrograms
> par(mfrow=c(1,2))
> plot(hc_single, main = "Single linkage clustering of variables", labels =
+       colnames(a))
> plot(hc_complete, main = "Complete linkage clustering of variables", labels
+       colnames(a))
> par(mfrow=c(1,1))
>
> clusters_single <- cutree(hc_single, k = 2)
>
> clusters_complete <- cutree(hc_complete, k = 2)
> clusters_single; clusters_complete
v1 v2 v3 v4 v5
 1  2  2  1  2
v1 v2 v3 v4 v5
 1  2  2  1  2
```

Interpretation:

Euclidean distance:

	v1	v2	v3	v4
v2	5.745843			
v3	5.146801	5.393380		
v4	4.642802	5.223147	7.157556	
v5	5.889456	4.907841	3.557968	6.472030

Single linkage:

```
hclust(d = dist.vars, method = "single")
```

```
Cluster method : single
Distance       : euclidean
```

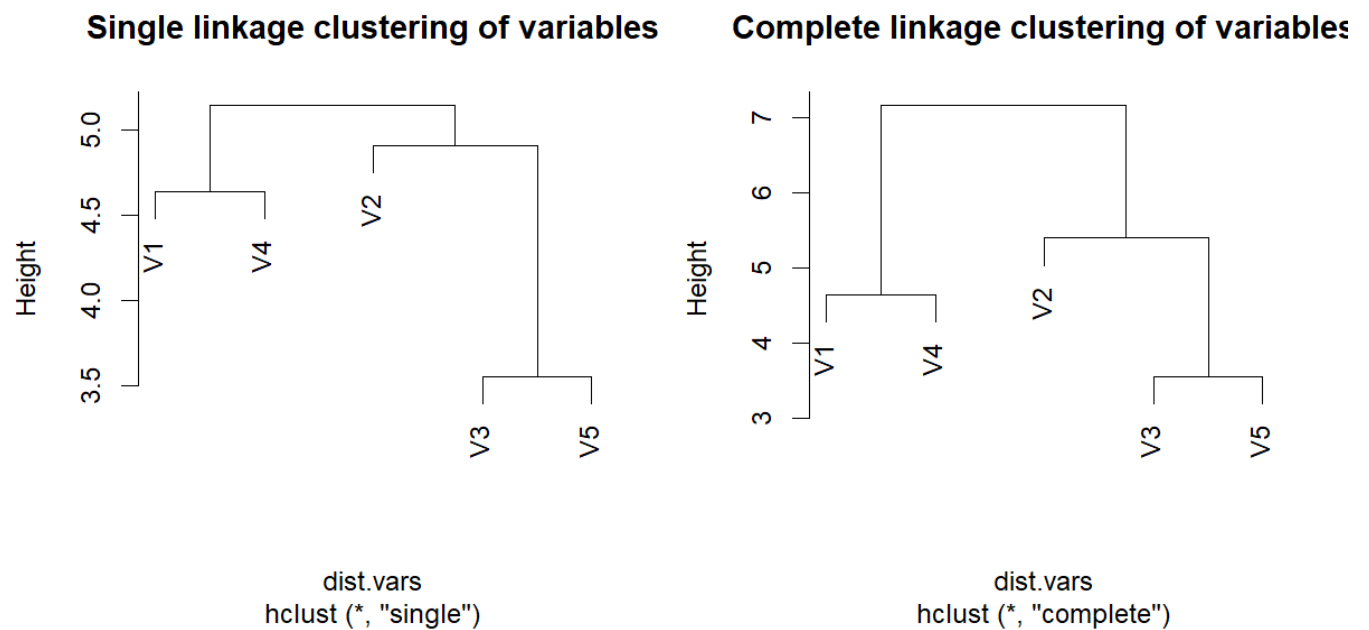
Number of objects: 5

Complete linkage:

```
call:  
hclust(d = dist.vars, method = "complete")
```

```
Cluster method   : complete  
Distance         : euclidean  
Number of objects: 5
```

The dendrograms:



Comparing the results:

Both linkage methods produced identical clusters, indicating a stable clustering structure. Variables V3 and V5 are very similar, while V1 and V4 form another group. Variable V2 joins later, showing weaker association.

Final Clusters

Cluster 1: {V1, V4}

Cluster 2: {V2, V3, V5}

