Techniques and Applications of Multivariate statistics (I)

Department of Statistics

Professor Yong-Seok Choi

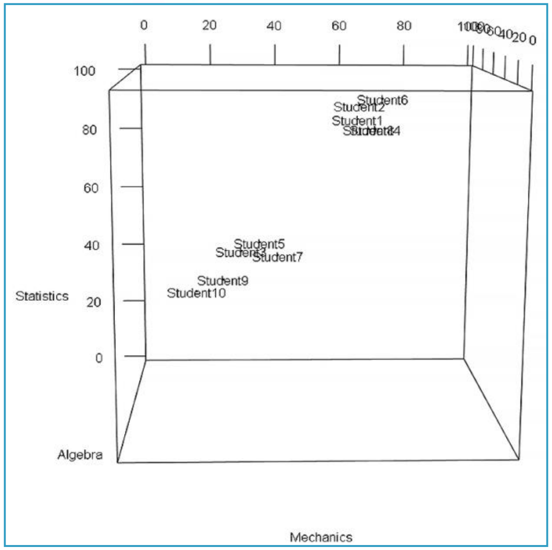
E-mail : yschoi@pusan.ac.kr

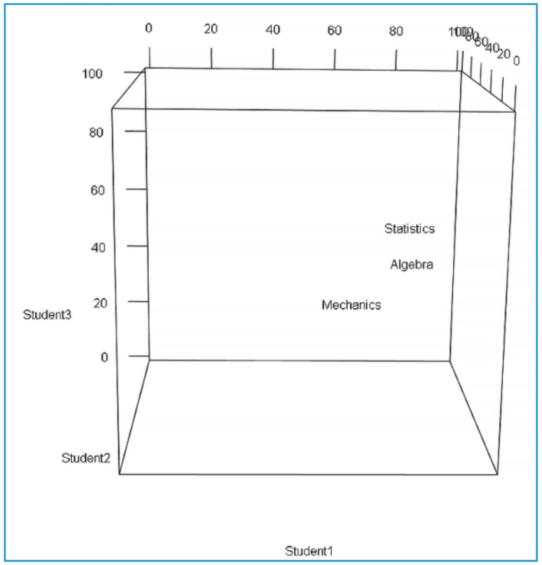
Home : yschoi.pusan.ac.kr

Contents

Multivariate Statistics (I) in Spring

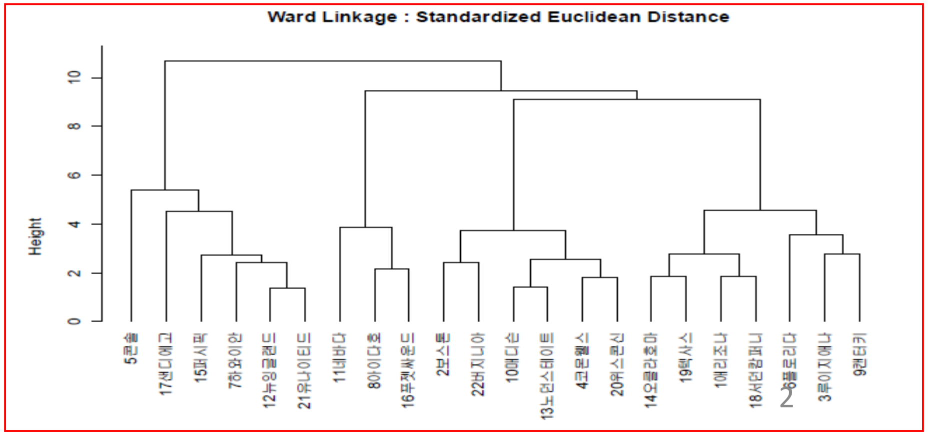
1. Multivariate Data Analysis
2. Principal Component Analysis (PCA)
3. Factor Analysis (FA)
4. Canonical Correlation Analysis (CCA)
5. Cluster Analysis (CA)





Multivariate Statistics (II) in Autumn

6. Discrimination and Classification Tree (DCT)
7. Multidimensional Scaling (MDS)
8. Correspondence Analysis (CRA)
9. Biplot
10. Shape Analysis



Multivariate Statistics (I)

1. Multivariate Data Analysis (MDA)

Contents

1.1 Multivariate data analysis

❖ Summarizing Data

[Example 1.1.1] 3 subjects' marks of 10 students

[Table 1.1.1] (3subjects.txt)

Students	Mechanics	Algebra	Statistics
Student1	65	85	85
Student2	65	80	90
Student3	30	40	50
Student4	70	83	82
Student5	35	43	52
Student6	72	82	92
Student7	40	43	48
Student8	68	83	82
Student9	25	32	43
Student10	17	51	35

[5] Stem-and-Leaf Plot

Mechanics	Algebra	Statistics
0 | 7	2 | 2	2 | 5
2 | 505	4 | 0331	4 | 3802
4 | 0	6 |	6 |
6 | 55802	8 | 02335	8 | 22502

[1] Descriptive Statistics

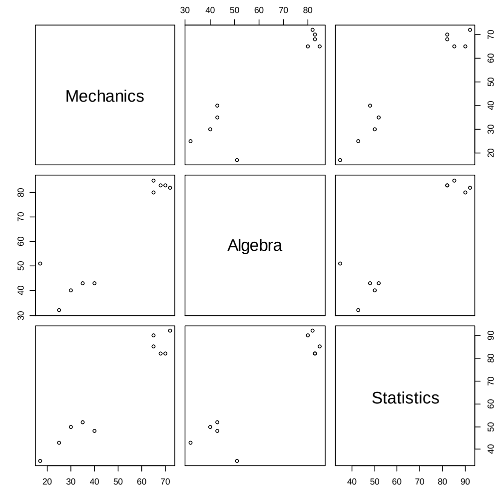
Mechanics	Algebra	Statistics
Min. :17.00	Min. :32.00	Min. :35.00
1st Qu. :31.25	1st Qu .:43.00	1st Qu. :48.50
Median :52.50	Median :65.50	Median :67.00
Mean :48.70	Mean :62.20	Mean :65.90
3rd Qu. :67.25	3rd Qu. :82.75	3rd Qu. :84.25
Max. :72.00	Max. :85.00	Max. :92.00

[2] Covariance Matrix

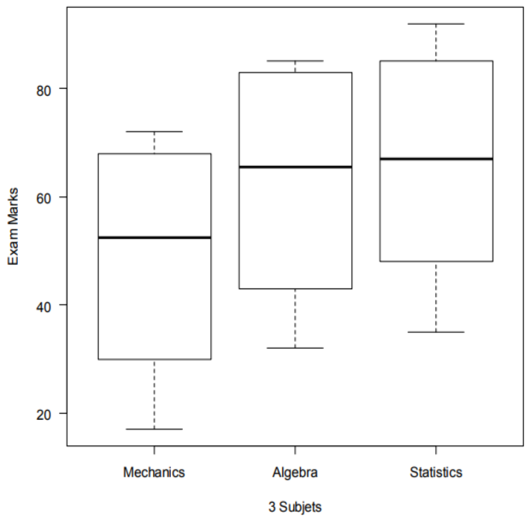
	Mechanics	Algebra	Statistics
Mechanics	453.3444	431.5111	459.0778
Algebra	431.5111	484.6222	450.2444
Statistics	459.0778	450.2444	487.8778

[3] Correlation Matrix

	Mechanics	Algebra	Statistics
Mechanics	1.0000000	0.9206110	0.9761501
Algebra	0.9206110	1.0000000	0.9259578
Statistics	0.9761501	0.9259578	1.0000000



[4] Multiple Scatter Plot

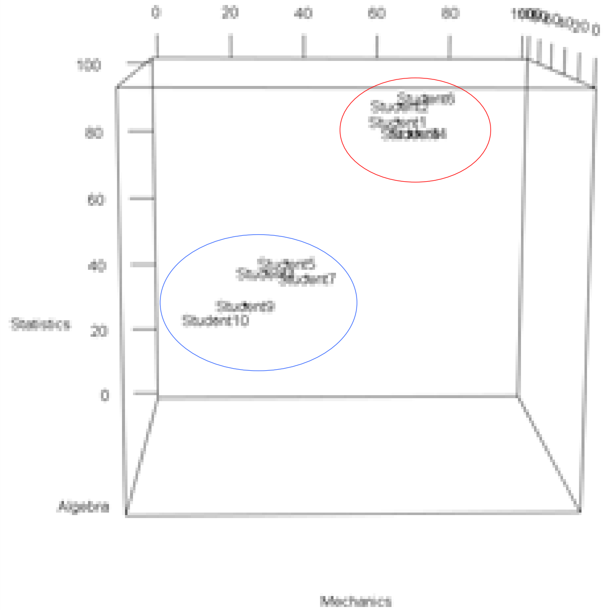


[5] Box Plot

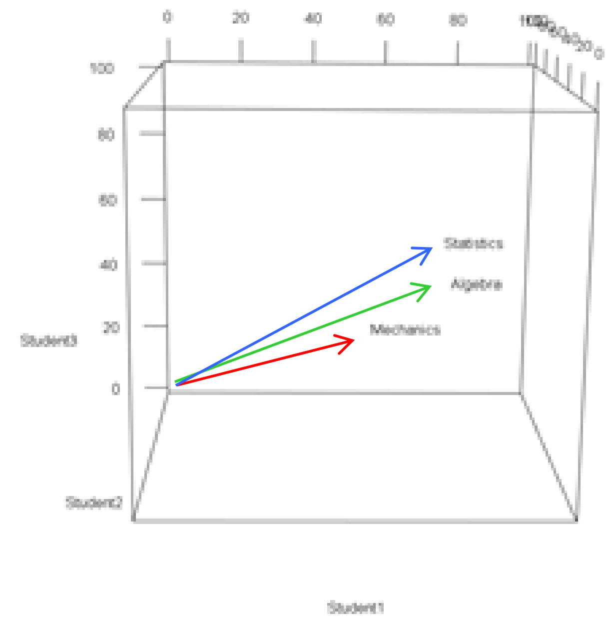
1.1 Multivariate data analysis

❖ Geometrical Representations of 3-dimensinal space

Students	Mechanics	Algebra	Statistics
Student1	65	85	85
Student2	65	80	90
Student3	30	40	50
Student4	70	83	82
Student5	35	43	52
Student6	72	82	92
Student7	40	43	48
Student8	68	83	82
Student9	25	32	43
Student10	17	51	35



a) n = 10 points in p–space



b) p = 3 points in n–space

[R-code 1.1.2] 3subjects-3d.R

```
Data1.1.1<-read.table("3subjects.txt", header=T)
X<-Data1.1.1
library(rgl)
```

```
# Observations in Variables Space
lim<-c(0, 100)
plot3d(X[,1], X[,2], X[,3],xlim=lim, ylim=lim, zlim=lim,
xlab="Mechanics", ylab="Algebra", zlab="Statistics")
text3d(X[,1], X[,2], X[,3],rownames(X))

# Variables in Observations Space
plot3d(X[1,], X[2,], X[3,], xlim=lim, ylim=lim, zlim=lim,
xlab="Student1", ylab="Student2", zlab="Student3")
text3d(X[1,], X[2,], X[3,], colnames(X))
```

1.2 Types of multivariate analysis techniques

- **Definition of Multivariate Analysis**
- A collection of techniques dealing with data containing observations on two or more variables.
- Multivariate data contain the *n observations* and *p variables*

❖ **Techniques based on the geometrical ideas**

✓ **R-Techniques** : Analyses based on the matrix of covariance or correlations between variables.
- PCA/FA/CCA/Biplot

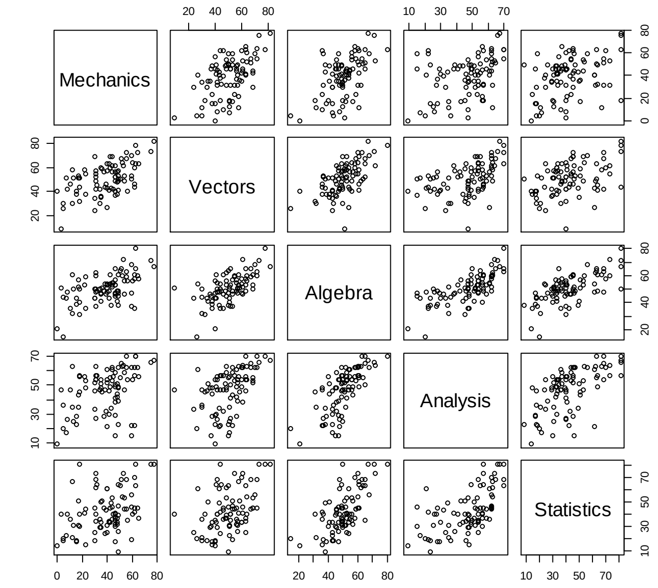
✓ **Q-Techniques** : Analyses based on the matrix of distances between observations.
- CA/DA/MDS/SCA/MCA/Biplot

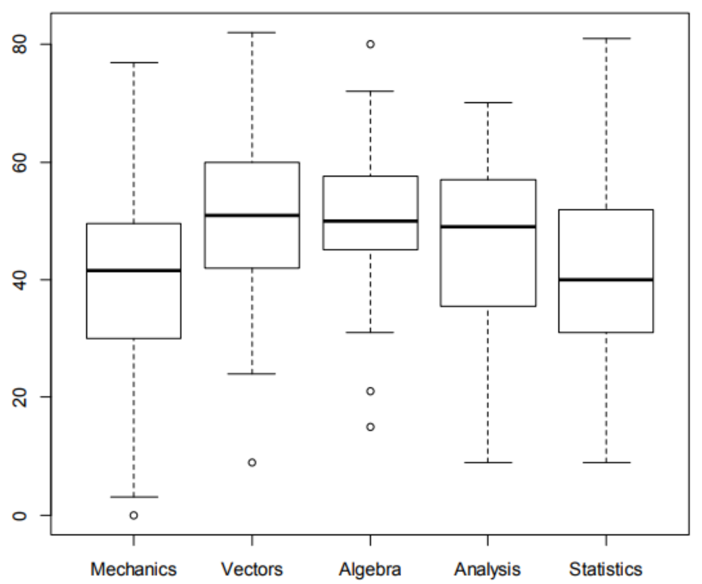
✓ **V-Techniques** : Visualization techniques based on relationships among variables or observations.
- Biplot/SCA/MCA/MDS/Stars Plot/Mosaic Plot

1.3 Introduction and visualization of multivariate data

- **[Data 1.3.1]** Examination marks on 5 subjects **(Mardia et al., 1979, pp. 3-4)**

	Closed-book		Open-book		
Student	Mechanics(c)	Vectors(c)	Algebra(o)	Analysis(o)	Statistics(o)
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
4	55	72	63	70	68
5	63	63	65	70	63
			⋮		
87	5	26	15	20	20
88	0	40	21	9	14





[R-code 1.3.1]
5subjects-boxsctter.R

```
Data1.3.1<-read.table("5subjects.txt", header=T)
X<-Data1.3.1[, -1]

# Multiple Scatter Plot
plot(X)

# Box Plot
boxplot(X)
```

Questions:

- How to combine or average these marks?
- Relationship between open-book and closed-book ?

Applications : PCA/FA/CCA

1.3 Introduction and visualization of multivariate data

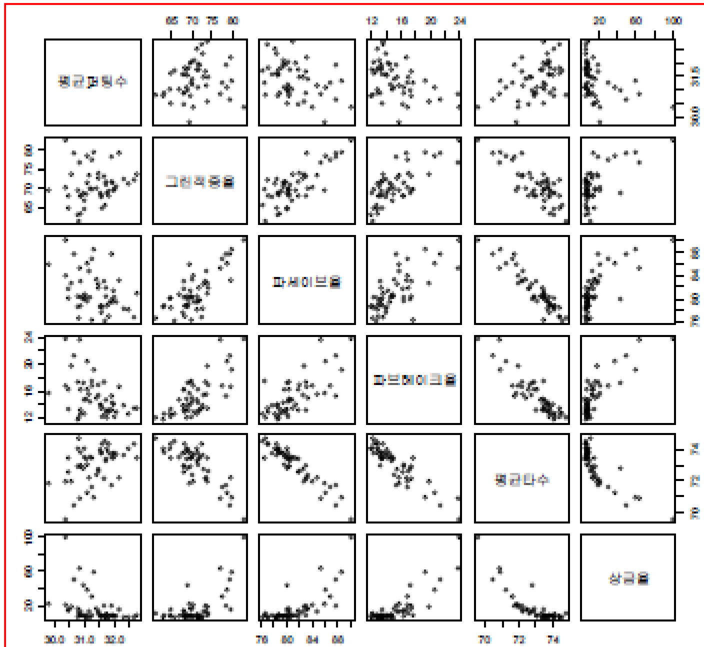
- **[Data 1.3.2] KLPGA player's grades (www.klpga.com, 2006)**

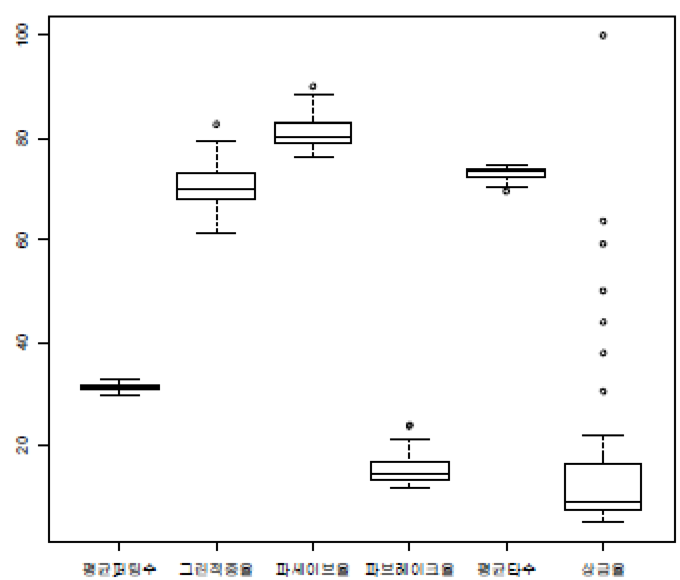
Putting average, Green in regulation % , Par save % , Par break % , Scoring average, Prize rate

선수	평균퍼팅수	그린적중률	파세이브율	파브레이크율	평균타수	상금율
1	30.36	82.72	90.12	23.77	69.58	100.0
2	30.85	76.94	85.29	23.69	70.85	63.7
3	31.33	79.63	88.52	19.26	70.93	59.4
4	30.64	79.32	87.65	21.30	70.47	50.2
5	30.97	68.86	79.97	17.17	72.79	44.0
6	31.09	78.59	86.03	20.48	71.23	38.0
7	31.25	77.47	86.88	16.82	71.61	30.5
8	29.81	69.53	85.94	15.63	71.84	21.8
9	30.36	70.37	83.95	16.82	71.97	21.1
10	31.91	78.28	87.71	15.32	71.73	19.2
⋮						
46	30.79	63.07	78.59	11.76	74.06	6.5
47	31.70	65.66	76.26	12.63	74.52	6.3
48	31.20	70.19	80.19	14.44	73.53	6.2
49	32.50	71.35	78.13	12.67	74.06	5.9
50	31.07	68.72	80.36	14.88	73.39	5.1

[R-code 1.3.2] klpga-boxscatter.R

```
Data1.3.2<-read.table("klpga.txt", header=T)
X<-Data1.3.2[, -1]

# Descriptive Statistics
summary(X)

# Covariance Matrix
cov(X)

# Correlation Matrix
cor(X)

# Multiple Scatter Plot
plot(X)

# Boxplot of 3 Subjects
boxplot(X)
```





Applications : CCA/CA/DCT

1.3 Introduction and visualization of multivariate data

- **[Data 1.3.4]** **Fisher's Iris flower data**(Johnson & Wichern, 2002, p. 657)

Iris Versicolor

Iris Setosa

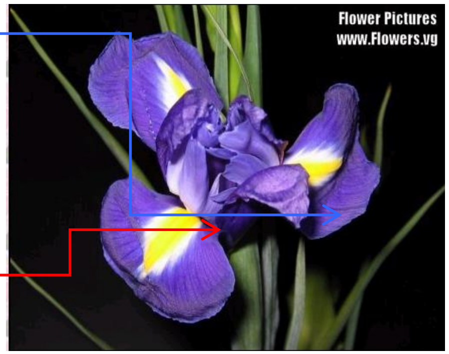
Iris Virginica

X1: Sepal length

X2: Sepal width

X3: Petal length

X4: Petal width



Setosa				Versicolor				Virginica			
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.8	3.0	1.4	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Questions:

– Ask to which species a new iris of unknown species belongs ?

– How to find the criteria for classifying?

Sir Ronald Aylmer Fisher
(17 February 1890 – 29 July 1962)
English statistician,
evolutionary biologist, and geneticist

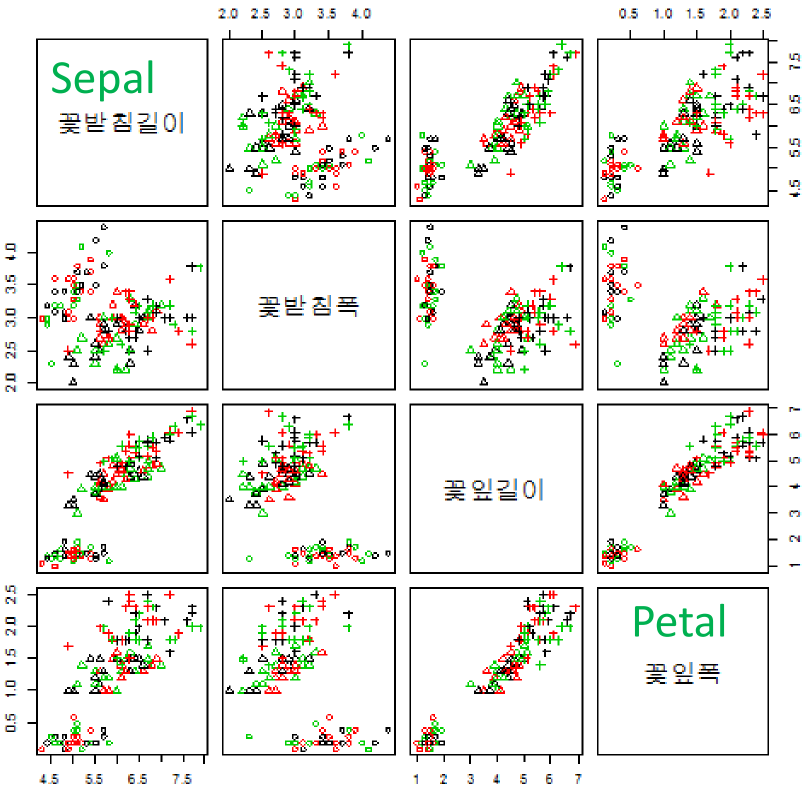
Applications : DCT/CA

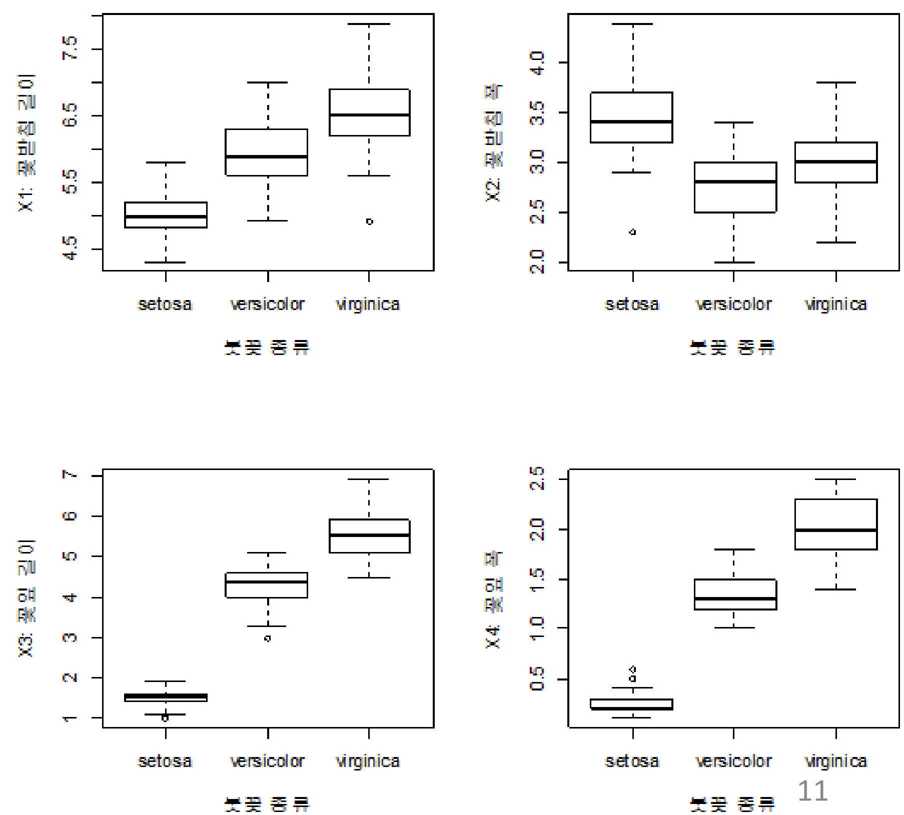
1.3 Introduction and visualization of multivariate data

- **Multiple Scatter Plot and Box Plot for** **Iris flower data**

[R-code 1.3.4] irisflower-boxscatter.R

```
Data1.3.4<-read.table("irisflower.txt", header=T)
X<-Data1.3.4[, -1]

# Box Plot
par(mfrow=c(2, 2))
boxplot(꽃받침길이~group, data=X, xlab="붓꽃 종류", ylab="X1: 꽃받침 길이")
boxplot(꽃받침폭~group, data=X, xlab="붓꽃 종류", ylab="X2: 꽃받침 폭")
boxplot(꽃잎길이~group, data=X, xlab="붓꽃 종류", ylab="X3: 꽃잎 길이")
boxplot(꽃잎폭~group, data=X, xlab="붓꽃 종류", ylab="X4: 꽃잎 폭")

# Multiple Scatter Plot
plot(X[, 1:4], pch=unclass(X[, 5]), col=1:3)
```



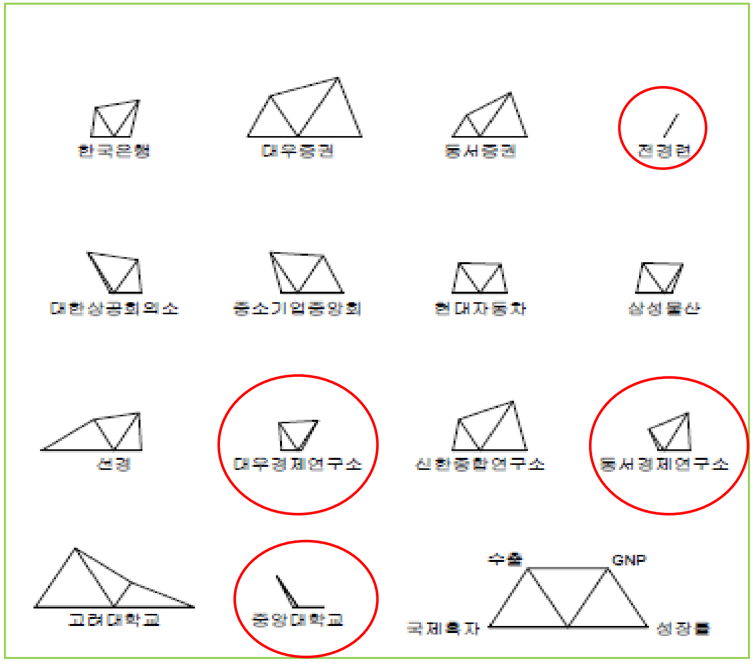


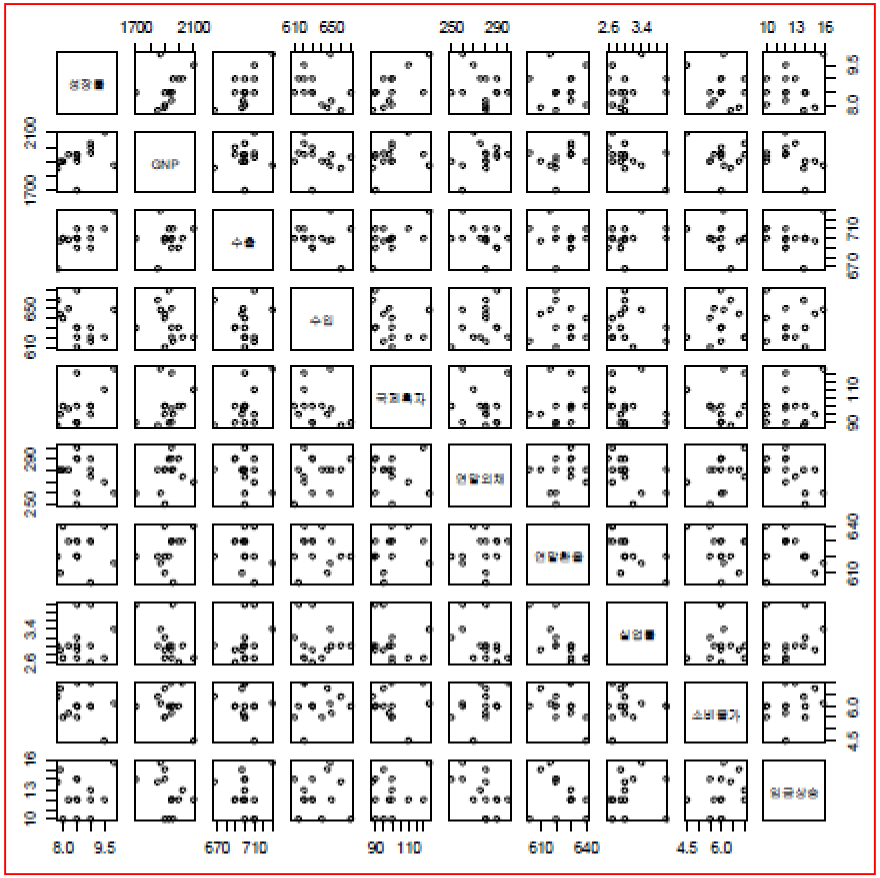
1.3 Introduction and visualization of multivariate data

❖ [Data 1.3.5] Economic Views Data of the 1990

기관	성장률	GNP	수출	수입	국제 흑자	연말 외채	연말 환율	실업률	소비 물가	임금 상승
한국은행	8.2	1950	698	650	98	280	630	3.0	5.7	12.0
대우증권	9.5	2100	710	620	110	270	640	2.7	4.5	12.0
동서증권	9.0	2000	690	630	100	290	630	2.6	6.0	12.0
전경련	7.8	1850	668	660	88	280	620	3.0	6.4	13.8
대한상공회의소	8.5	1928	710	670	90	290	620	3.0	6.0	10.0
중소기업중앙회	9.0	1958	710	615	95	280	603	4.0	6.0	10.0
현대자동차	8.5	1900	700	610	100	250	620	3.2	5.5	14.0
삼성물산	8.0	1900	700	640	100	280	640	2.7	5.5	10.0
선경	8.5	1950	700	620	120	300	630	2.7	7.0	12.0
대우경제연구소	7.9	1900	697	645	95	280	610	2.9	6.8	15.0
신한종합연구소	9.0	2030	700	620	100	275	630	3.0	7.0	13.0
동서경제연구소	8.5	1950	690	630	90	290	630	2.9	6.0	12.0
고려대학교	9.9	1870	729	649	123	260	616	3.4	6.1	15.8
중앙대학교	8.5	1700	700	630	90	260	620	4.0	6.0	14.0

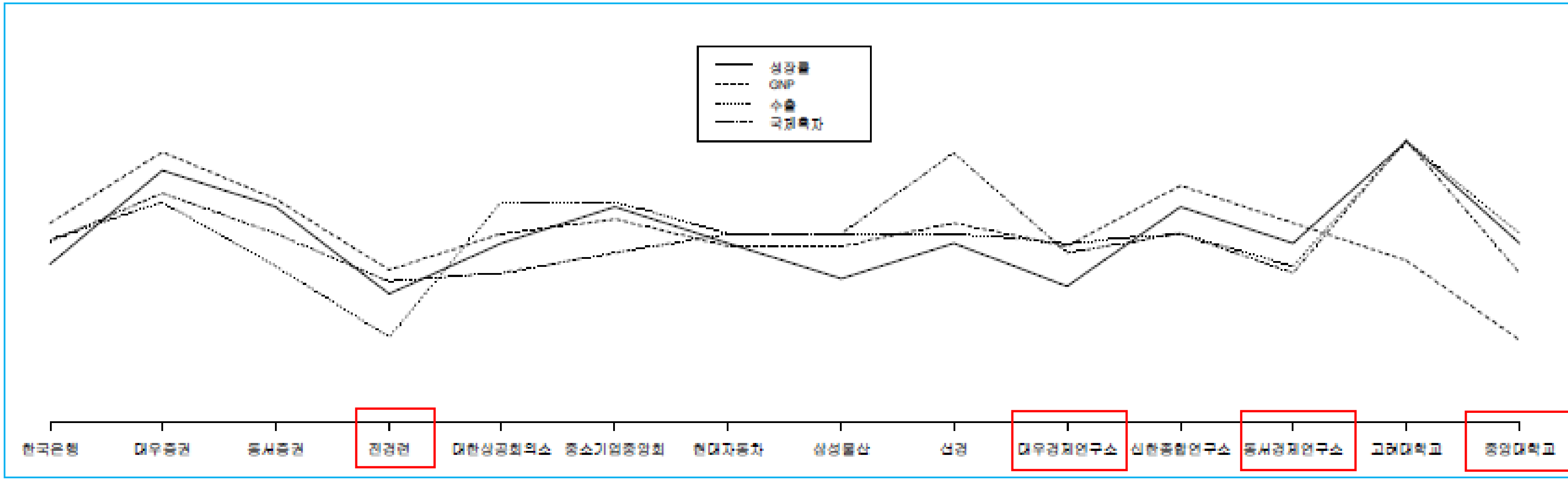
Growth /Export/Import/ Black-ink balance/ Foreign debt /Exchange rate/ Unemployed rate /Consumer price rate /Wage increase rate





1.3 Introduction and visualization of multivariate data

❖ [Data 1.3.5] Economic Views Data



[R-code 1.3.5] economicview-starscatterparcoord.R

```
Data1.3.5<-read.table("economicview.txt", header=T)
X<-Data1.3.5[, -1]

# Multiple Scatter Plot
plot(X)

# Star Plot
X<-scale(as.matrix(X[, c(1,2,3,5)]))
rownames(X)<-Data1.3.5[, 1]
stars(X,key.loc=c(8,2), full = FALSE)

# Parallel Coordinate Plot
library(gclus)
parcoordlabel<-function(x, col = 1, lty = 1,  var.label=F,...)
{
     rx <-lapply(X, range, na.rm = TRUE)
    matplot(1L:ncol(x), t(x), type = "l", col = col, lty = 1:nrow(X), lwd=1.5,
        xlab = "", ylab = "", axes = FALSE,
           ylim=c(-4, 4), xlim=c(1, nrow(X)), ...)
    axis(1, at = 1L:ncol(x), labels = colnames(x))
    legend("top", horiz=F, legend=colnames(X), lty=1:nrow(X),
          col=1, cex=0.8, lwd=1.5)
    for (i in 1L:ncol(x))
      invisible()
 }
windows(height=5, width=12)
parcoordlabel(t(X))
```

1.4 Matrix representation and descriptive statistics of multivariate data

❖ Representation

✓ $n \times p$ data matrix

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = (x_{ij}),\ i = 1, \ldots, n;\ j = 1, \ldots, p$$

$$X = \begin{bmatrix} 65 & 85 & 85 \\ 65 & 80 & 90 \\ 30 & 40 & 50 \\ 70 & 83 & 82 \\ 35 & 43 & 52 \\ 72 & 82 & 92 \\ 40 & 43 & 48 \\ 68 & 83 & 82 \\ 25 & 32 & 43 \\ 17 & 51 & 35 \end{bmatrix}$$

10x3 Data Matrix from [Data 1.1.1]

$$X = \begin{bmatrix} x_1^t \\ \vdots \\ x_i^t \\ \vdots \\ x_n^t \end{bmatrix} = [x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)}]$$

✓ Centred data matrix : $Y = X - \frac{1}{n}JX = HX$

✓ Standardized data matrix : $Z = HXD_s^{-1/2}$

where $H = I - \frac{1}{n}J$: centring matrix and symmetric idempotent

$J =$ Square matrix with all elements 1, $I =$ unit matrix

$D_s^{1/2} = diag(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$: SD matrix

Representation of Data in Space:

$$x_i \in \mathbb{Q}^p,\ i = 1, \ldots, n$$
$$x_{(j)} \in \mathbb{R}^n,\ j = 1, \ldots, p$$

I) Sample Summary Statistics

- (Sample) mean vector $\bar{x} = [\bar{x}_1, \ldots, \bar{x}_j, \ldots, \bar{x}_p]^t = \dfrac{X^t 1_n}{n}$

- (Sample variance-) covariance matrix

$$S = (s_{kj}) = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ & & & \vdots & & \vdots \\ & & & s_{jj} & \cdots & s_{jp} \\ & (\text{대칭}) & & & & \\ & & & & s_{pp} & \end{bmatrix}$$

$$s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

$$s_{kj} = s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j)$$

- Algebraic relationship : $S = \dfrac{1}{n-1} Y^t Y = \dfrac{1}{n-1} X^t H X.$

1.4 Matrix representation and descriptive statistics of multivariate data

- Correlation matrix

$$\boldsymbol{y}_{(j)} = (y_{1j}, \ldots, y_{nj})^t = (x_{1j} - \overline{x}_j, \ldots, x_{nj} - \overline{x}_j)^t, \quad j = 1, \cdots, p$$

$$r_{kj} = \frac{s_{kj}}{\sqrt{s_{kk}}\sqrt{s_{jj}}} = \frac{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(x_{ij} - \overline{x}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} \quad \text{if } k \neq j$$

$$R = (r_{kj}) = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ & & & \vdots & & \vdots \\ & & & 1 & \cdots & r_{jp} \\ & (\text{대칭}) & & & & \vdots \\ & & & & & 1 \end{bmatrix}$$

$$\cos\theta_{kj} = \frac{\boldsymbol{y}_{(k)}^t \boldsymbol{y}_{(j)}}{\|\boldsymbol{y}_{(k)}\| \, \|\boldsymbol{y}_{(j)}\|} = \frac{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(x_{ij} - \overline{x}_j)}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}} = r_{kj}$$

- Algebraic relationship :

$\boxed{R = D_s^{-1/2} S D_s^{-1/2}}$ where $D_s^{-1/2} = (D_s^{1/2})^{-1}$, $D_s^{1/2} = diag(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$: standard deviation matrix (squared root matrix)

$$\boxed{R = \frac{1}{n-1} Z^t Z}$$

Note :

The measurement units of variables are different or some variables have widely differing variances => correlation matrix R

- Measures for amount of variation from the centroid:

- generalized variance = $|S|$ *or* $|R|$
- total variance = $tr(S) = s_{11} + \cdots + s_{pp}$, $or\ tr(R)$

Notes:

- Large values indicate a high degree of scatter about mean vector and low values represent concentration about mean vector
- |S|=0 : collinearity among variables
- |R|=0 : correlation among variables
- |S| or |R| plays an important role in MLE/FA(Sec 3.4 MLFA)/DCT
- tr(S) or tr(R) is a useful concept in PCA(Sec 2.4)/FA(Sec 3.4) : goodness-of-fit

1.4 Matrix representation and descriptive statistics of multivariate data

- **[Example 1.4.1]** **S and R for examination marks on 3 subjects**

$\bar{\mathbf{x}}$ = (48.70, 62.20, 65.90)

- **Results of** [R-code 1.4.1] 3subjects-covcorr.R

자료행렬 X				중심화 자료행렬 Y				표준화 자료행렬 Z			
	Mechanics	Algebra	Statistics		Mechanics	Algebra	Statistics		Mechanics	Algebra	Statistics
1	65	85	85	[1,]	16.3	22.8	19.1	[1,]	0.7655498	1.0356981	0.8647247
2	65	80	90	[2,]	16.3	17.8	24.1	[2,]	0.7655498	0.8085713	1.0910924
3	30	40	50	[3,]	-18.7	-22.2	-15.9	[3,]	-0.8782688	-1.0084429	-0.7198493
4	70	83	82	[4,]	21.3	20.8	16.1	[4,]	1.0003810	0.9448474	0.7289040
5	35	43	52	[5,]	-13.7	-19.2	-13.9	[5,]	-0.6434376	-0.8721668	-0.6293022
6	72	82	92	[6,]	23.3	19.8	26.1	[6,]	1.0943135	0.8994220	1.1816395
7	40	43	48	[7,]	-8.7	-19.2	-17.9	[7,]	-0.4086063	-0.8721668	-0.8103964
8	68	83	82	[8,]	19.3	20.8	16.1	[8,]	0.9064486	0.9448474	0.7289040
9	25	32	43	[9,]	-23.7	-30.2	-22.9	[9,]	-1.1131000	-1.3718457	-1.0367641
10	17	51	35	[10,]	-31.7	-11.2	-30.9	[10	-1.4888300	-0.5087640	-1.3989525

공분산행렬 S				상관행렬 R			
	Mechanics	Algebra	Statistics		Mechanics	Algebra	Statistics
Mechanics	453.3444	431.5111	459.0778	Mechanics	1.0000000	0.9206110	0.9761501
Algebra	431.5111	484.6222	450.2444	Algebra	0.9206110	1.0000000	0.9259578
Statistics	459.0778	450.2444	487.8778	Statistics	0.9761501	0.9259578	1.0000000

변동량척도	
일반화분산 $\lvert S \rvert = 690375.8$	일반화분산 $\lvert R \rvert = 0.006440846$
총분산 $tr(S) = 1425.844$	총분산 $tr(R) = 3$

1.4 Matrix representation and descriptive statistics of multivariate data

- **[Example 1.4.2]** **S and R for iris flower data**
- **Results of** [R-code 1.4.2] irisflower-covcorr.R

	S				R				$\lvert S \rvert$	$tr(S)$	$\lvert R \rvert$	$tr(R)$
setosa	0.124	0.099	0.016	0.010	1.000	0.743	0.267	0.278	0.000002	0.309204	0.353359	4.000000
	0.099	0.144	0.012	0.009	0.743	1.000	0.178	0.233				
	0.016	0.012	0.030	0.006	0.267	0.178	1.000	0.332				
	0.010	0.009	0.006	0.011	0.278	0.233	0.332	1.000				
versicolor	0.266	0.085	0.183	0.056	1.000	0.526	0.754	0.546	0.000019	0.624824	0.083594	4.000000
	0.085	0.098	0.083	0.041	0.526	1.000	0.561	0.664				
	0.183	0.083	0.221	0.073	0.754	0.561	1.000	0.787				
	0.056	0.041	0.073	0.039	0.546	0.664	0.787	1.000				
virginica	0.404	0.094	0.303	0.049	1.000	0.457	0.864	0.281	0.000133	0.888367	0.137390	4.000000
	0.094	0.104	0.071	0.048	0.457	1.000	0.401	0.538				
	0.303	0.071	0.305	0.049	0.864	0.401	1.000	0.322				
	0.049	0.048	0.049	0.075	0.281	0.538	0.322	1.000				

1.5 Multivariate data distance

(1) Euclidean Distance

$$d_{rs} = \|\boldsymbol{x}_r - \boldsymbol{x}_s\| = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^t(\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2} = \left[\sum_{j=1}^{p}(x_{rj} - x_{sj})^2\right]^{1/2}$$

(2) Weighted Euclidean Distance

$$d_{rs} = \|\boldsymbol{x}_r - \boldsymbol{x}_s\|_{D_w} = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^t D_w (\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2} = \left[\sum_{j=1}^{p} w_j (x_{rj} - x_{sj})^2\right]^{\frac{1}{2}}$$

(3) Nomalized Eucliean Distance

$$d_r = \|\boldsymbol{x}_r - \boldsymbol{x}_s\|_{D_s^{-1}} = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^t D_s^{-1} (\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2} = \left[\sum_{j=1}^{p} \frac{1}{s_{jj}} (x_{rj} - x_{sj})^2\right]^{\frac{1}{2}}$$

(4) Mahalanobis Distance

$$d_r = \|\boldsymbol{x}_r - \boldsymbol{x}_s\|_{S^{-1}} = \left[(\boldsymbol{x}_r - \boldsymbol{x}_s)^t S^{-1} (\boldsymbol{x}_r - \boldsymbol{x}_s)\right]^{1/2}$$

(5) City-Block Distance

$$d_{rs} = \sum_{j=1}^{p} |x_{rj} - x_{sj}|$$

(6) Minkowski Distance

$$d_{rs} = \left[\sum_{j=1}^{p} w_j |x_{rj} - x_{sj}|^m\right]^{1/m}, \quad m \geq 1$$

Hermann Minkowski (1864–1909): Russian mathematician
His former student Albert Einstein's special theory of relativity (1905) could be understood geometrically as a theory of four-dimensional space–time since known as the "Minkowski spacetime"..

1.5 Multivariate data distance

- **[Example 1.5.1][DATA 1.1.1]**

Result of [R-code 1.5.1] 3subjects-distances.R

```
Data1.1.1<-read.table("3subjects.txt", header=T)
X<-Data1.1.1[,-1]
X<-as.matrix(Data1.1.1[,-1])

n<-nrow(X)
xbar<-t(X)%*%matrix(1,n,1)/n  # 평균벡터
I<-diag(n)
J<-matrix(1,n,n)
H<-I-1/n*J                     # 중심화행렬
Y<-H%*%X                       # 중심화 자료행렬
S<-t(Y)%*%Y/(n-1)              # 공분산행렬
D<-diag(1/sqrt(diag(S)))       # 표준편차행렬의 역
Z<-Y%*%D                       # 표준화자료행렬
colnames(Z)<-colnames(X)

# 유클리드 거리
de <- as.matrix(dist(X, method="euclidean"))
de <- as.dist(de)
round(de, 3)
```

```
# 표준화 유클리드 거리
ds <- as.matrix(dist(Z, method="euclidean"))
ds <- as.dist(ds)
round(ds, 3)

# 마할라노비스 거리
library(biotools)
dm<-D2.dist(X, S)
round(sqrt(dm), 3)

# 시티블럭 거리
dc <- as.matrix(dist(X, method="manhattan"))
dc <- as.dist(dc)
round(dc, 3)
```

Euclidean distance

$$x_1 - x_2 = (65-65,\ 85-80,\ 85-90)^t = (0,\ 5,\ -5)^t$$

	1	2	3	4	5	6	7	8	9
2	7.071								
3	66.895	66.521							
4	6.164	9.899	66.880						
5	61.262	60.934	6.164	61.033					
6	10.344	7.550	72.746	10.247	67.007				
7	61.303	61.303	10.630	60.465	6.403	66.940			
8	4.690	9.055	65.704	2.000	59.908	10.817	59.498		
9	78.568	78.186	11.747	78.403	17.378	84.321	19.261	77.272	
10	77.201	78.549	22.694	77.730	26.019	85.059	27.604	76.381	22.113

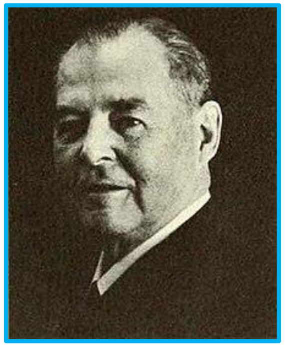
$$\begin{aligned} d_{12} &= \left[(x_1-x_2)^t(x_1-x_2)\right]^{1/2} \\ &= \left[(0,\ 5,\ -5)(0,\ 5,\ -5)^t\right]^{1/2} \\ &= \left[0^2+5^2+(-5)^2\right]^{1/2} = 7.071 \end{aligned}$$

1.6 Multivariate normal distribution and its useful properties

- **Univariate Distributions vs. Multivariate Distributions**

Univariate	Multivariate
Normal dist.	Multivariate normal dist.
Chi-square dist.	Wishart dist.
t-dist.	Hotelling's T^2 dist.
F-dist.	Wilks' Λ dist.

Harold Hotelling(September 29, 1895 – December 26, 1973) was an American mathematical statistician and an influential economic theorist, Hotelling's T-squared distribution in statistics. He also developed and named the principal component analysis method widely used.



John Wishart(28 November 1898 – 14 July 1956) was a Scottish mathematician and agricultural statistician. worked successively at University College London with Karl Pearson, at Rothamsted Experimental Station with Ronald Fisher. He first formulated the Wishart distribution in his honour, in 1928 .

- **Univariate normal distribution :** $X \sim N(\mu, \ \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty$$

Samuel Stanley Wilks (June 17, 1906 – March 7, 1964) was an American mathematician and academic who played an important role in the development of mathematical statistics. He provided Wilks' Theorem in the theory of likelihood ratio tests, where he showed the distribution of log likelihood ratios is asymptotically χ 2 .

1.6 Multivariate normal distribution and its useful properties

- **Multivariate Normal Distribution:** $N_p(\boldsymbol{\mu},\ \Sigma)$

$$f(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

Likelihood function： $L(\boldsymbol{\mu},\ \Sigma) = \prod_{i=1}^{n} f(\boldsymbol{x_i})$

$$= \frac{1}{(\sqrt{2\pi})^{np} |\Sigma|^{n/2}} exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x_i}-\boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x_i}-\boldsymbol{\mu})\right)$$

❖ Bivariate Normal Distribution: $N_2(\boldsymbol{\mu},\ \Sigma)$

$$f(x_1,\ x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}} \times \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

1.6 Multivariate normal distribution and its useful properties

- **Deriving** $N_2(\boldsymbol{\mu}, \Sigma)$ **from** $N_p(\boldsymbol{\mu}, \Sigma)$

$$\boldsymbol{x} = (x_1, x_2)' \sim N_2(\boldsymbol{\mu}, \Sigma)$$

$$\textbf{with} \quad \boldsymbol{\mu} = (\mu_1, \mu_2)' \quad \textbf{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} \quad : \text{correlation coefficient}$$

$$|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} = \sigma_{11}\sigma_{22}(1-\rho_{12}^2)$$

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} \sigma_{22} & -\sigma_{21} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

$$= \frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}\begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix}$$

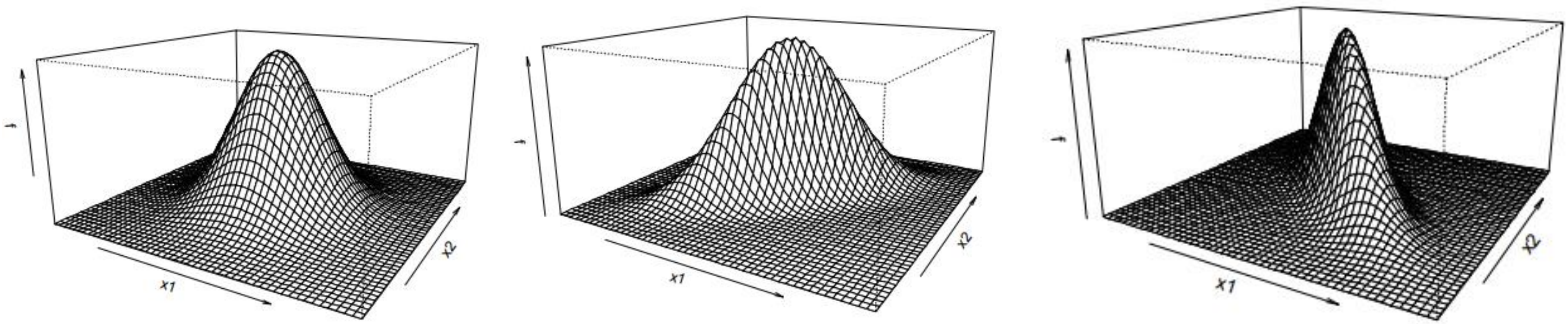
$$(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = [x_1-\mu_1, x_2-\mu_2]\ \frac{1}{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}$$

$$\times \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}$$

$$= \frac{1}{1-\rho_{12}^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

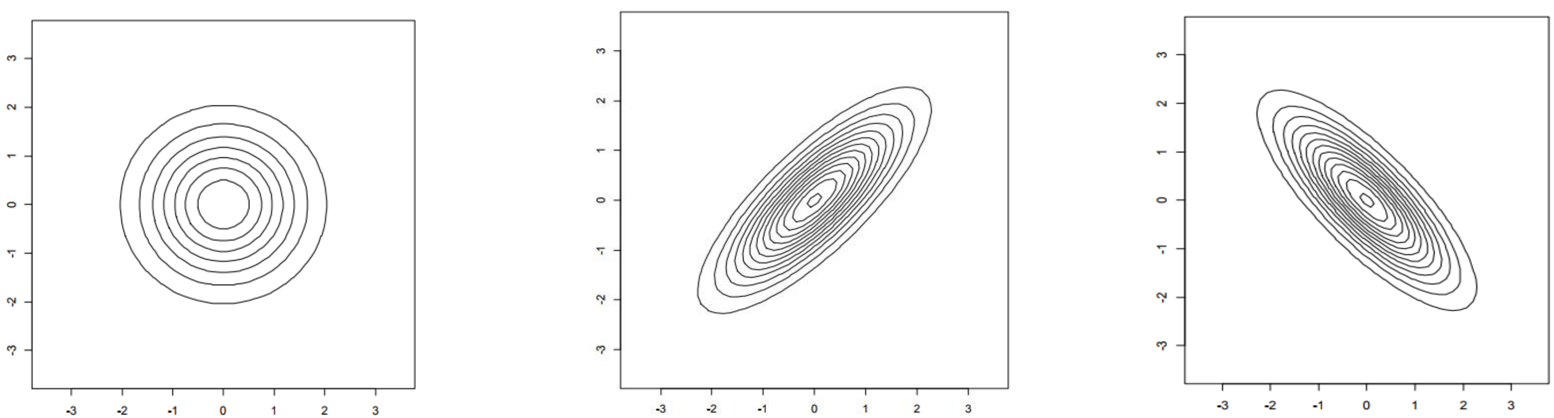
1.6 Multivariate normal distribution and its useful properties

- **Bivariate normal distribution** $\mu = (0.0,\ 0.0)^t$



(a) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = 0.0$ (b) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = 0.8$ (c) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = -0.8$

- **Contour of normal distribution** $\{\boldsymbol{x} : \boldsymbol{x}^t \Sigma^{-1} \boldsymbol{x} = c^2\}$: Ellipsoid



(a) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = 0.0$ (b) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = 0.8$ (c) $\sigma_{11} = \sigma_{22} = 1,\ \rho_{12} = -0.8$

1.7 Wishart W-dist and Hotelling's T^2-dist

Karl Pearson(27 March 1857 – 27 April 1936) : English mathematician and biostatistician

• Chi-square dist. and Wishart dist.

■ Chi-square distribution in the univariate case

$$- \ p=1, \quad x_1, \dots, x_n \overset{iid}{\sim} N(\mu, \ \sigma^2) \Rightarrow z_i = \frac{x_i - \mu}{\sigma}, i=1,\dots,n, \quad \overset{iid}{\sim} N(0, \ 1).$$

$$\Rightarrow \sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2} \sim \chi_n^2$$

$$\Leftrightarrow w = \sum_{i=1}^{n} (x_i-\mu)^2 \sim \sigma^2 \chi_n^2$$

Using $\bar{x}$ instead of $\mu \Rightarrow$ $\boxed{w = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s^2 \sim \sigma^2 \chi_{n-1}^2}$

Friedrich Robert Helmert (July 31, 1843 – June 15, 1917): German geodesist

■ Wishart distribution in the multivariate case

Let $\boldsymbol{x}_1 = (x_{11}, \dots, x_{1p})', \dots, \boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})', \dots, \boldsymbol{x}_n = (x_{n1}, \dots, x_{np})' \sim N_p(\boldsymbol{\mu}, \Sigma)$

$$W = \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})' \sim W_p(n, \Sigma)$$: Wishart distribution with n df

Using $\bar{\boldsymbol{x}}$ instead of $\boldsymbol{\mu} \Rightarrow$ $\boxed{W = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' = (n-1)S = V \sim W_p(n-1, \Sigma)}$

This distribution known as the Helmert distribution was first described by the German statistician Friedrich Robert Helmert in 1875–6. This was independently rediscovered by the English mathematician Karl Pearson in goodness of fit, for which he developed his chi-squared test, published in 1900 with computed table.

1.7 Wishart W-dist and Hotelling's T^2-dist

- **t-dist. and T^2-dist.**

William Sealy Gosset (13 June 1876 – 16 October 1937) was an English statistician. He published under the pen name **Student**, and developed the Student's t-distribution. Guinness prohibited its employees from publishing any papers regardless of the contained information.

  - t-distribution in the univariate case

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \longrightarrow t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

$$\boxed{t^2 = \frac{(\bar{x} - \mu)^2}{(s/\sqrt{n})^2} = n(\bar{x} - \mu)(s^2)^{-1}(\bar{x} - \mu)}$$

  - T^2-distribution in the multivariate case

$$\boxed{T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' S^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim T_p^2(n-1) = \frac{(n-1)p}{n-p} F_{p,\, n-p}}$$

Note : $p = 1 : F_{1,\ m} = T_1^2(m)$

$$- \ t \sim t(m) \quad \Longrightarrow \quad t^2 \sim F_{1,\ m} \sim T_1^2(m)$$

1.8 Testing multivariate normality

- **Steps for chi-squire plot**

Prasanta Chandra Mahalanobis (29 June 1893 – 28 June 1972): an Indian scientist and applied statistician

Mahalanobis distance: $\boxed{m_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^t S^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}),\ i = 1,\ \dots,\ n}$

[Step 1] Order the Mahalanobis distances from smallest to largest as

$$\boxed{m_{(1)}^2 \le m_{(2)}^2 \le \cdots \le m_{(n)}^2}$$

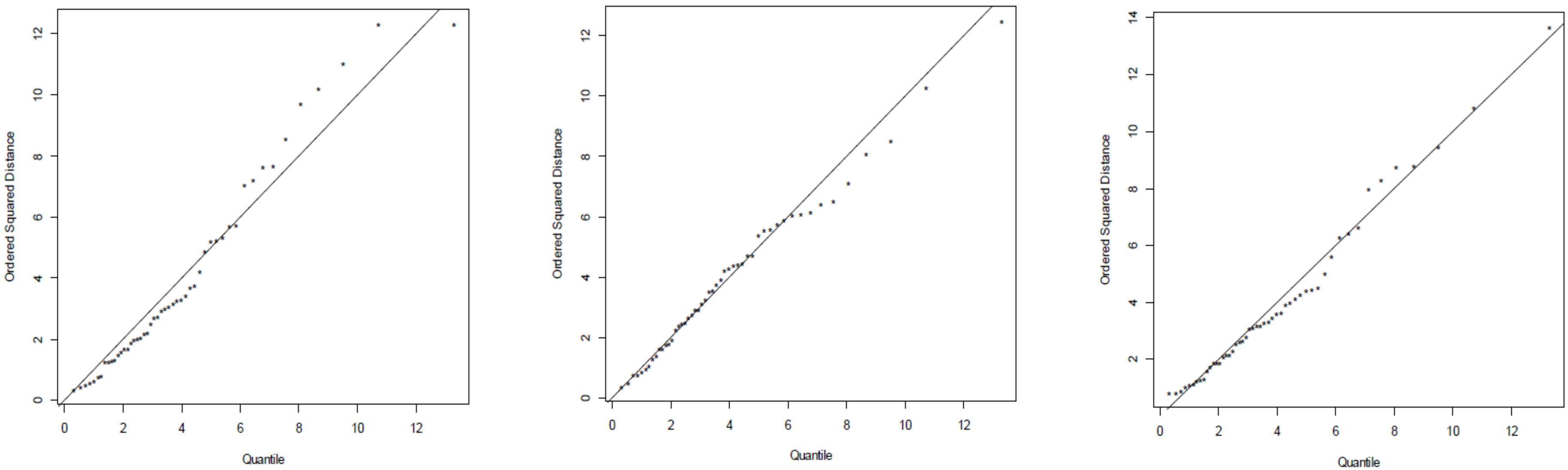
[Step 2] Calculate the $100\left(i - \frac{1}{2}\right)/n,\ i = 1, \dots, n$ percentile $q_{c(i)}$ of chi-square distribution with *p* d.f.

[Step 3] Plot the pairs $(q_{c(i)},\ m_{(i)}^2),\ i = 1,\ \dots,\ n$

[Step 4] Check the straightness of plot.

[Example1.8.1] Normality test: Chi-square Plot

$$r_Q = \frac{\sum_{i=1}^{n}(q_{c(i)} - \overline{q})(m_{(i)}^2 - \overline{m})}{\sqrt{\sum_{i=1}^{n}(q_{c(i)} - \overline{q})^2}\sqrt{\sum_{i=1}^{n}(m_{(i)}^2 - \overline{m})^2}}$$



$r_Q = 0.9863187$ (a) Setosa

$r_Q = 0.9943037$ (b) Versicolor

$r_Q = 0.9927457$ (c) Virginica

[Figure 1.8.1] Chi-square plot and correlation coefficient of iris data

[Example1.8.2] Normality test : Multivaiate Skewness & Kurtosis

Kantilal Vardichand Mardia (born 1935) : Indian statistician specializing in multivariate analysis and statistical shape analysis

$$m_{rs} = (x_r - \bar{x})^t S^{-1} (x_s - \bar{x}), \quad r, s = 1, \ldots, n$$

$$b_{1p} = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} m_{rs}^3 \simeq \frac{6}{n} \chi^2_{(p(p+1)(p+2)/6)}$$

$$b_{2p} = \frac{1}{n} \sum_{r=1}^{n} m_{rs}^2 \simeq N\left(p(p+2), \frac{8p(p+2)}{n}\right)$$

$\boxed{n < 20}$ $$b_{1p} = \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} m_{rs}^3 \sim \frac{6}{nk} \chi^2_{(p(p+1)(p+2)/6)}$$

$$k = (p+1)(n+1)(n+3) / (n(n+1)(p+1) - 6)$$

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : setosa

   g1p             : 3.079721
   chi.skew        : 25.66434
   p.value.skew    : 0.1771859

   g2p             : 26.53766
   z.kurtosis      : 1.294992
   p.value.kurt    : 0.1953229

   chi.small.skew  : 27.85973
   p.value.small   : 0.1127617

   Result    : Data are multivariate normal.
---------------------------------------
```

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : versicolor

   g1p             : 3.022201
   chi.skew        : 25.18501
   p.value.skew    : 0.1944445

   g2p             : 22.87938
   z.kurtosis      : -0.5718664
   p.value.kurt    : 0.5674125

   chi.small.skew  : 27.33939
   p.value.small   : 0.1259826

   Result    : Data are multivariate normal.
---------------------------------------
```

```
Mardia's Multivariate Normality Test
---------------------------------------
   data : virginica

   g1p             : 3.152472
   chi.skew        : 26.2706
   p.value.skew    : 0.1570597

   g2p             : 24.29906
   z.kurtosis      : 0.1526142
   p.value.kurt    : 0.8787025

   chi.small.skew  : 28.51784
   p.value.small   : 0.09769648

   Result    : Data are multivariate normal.
---------------------------------------
```


[Figure 1.8.2] MVN of iris data ← [R-code 1.8.2](iris-MVNtest.R)

1.9 R for EDA : Practice Time

R-code:

EDA	
summary()	Descriptive Statistics
cov(), cor()	Covariance, Correlation Matrices
plot() boxplot() stem()	Multiple Scatter Plot Multiple Box-and-Whisker Plot Stem-and-Leaf Plot
star() parcoordlabel()	Stars Plot Parallel Coordinate Plot, library(gclus)
mardiaTest() mvn()	Skewness & Kurtosis Tests, library(MVN) qqplot=TRUE mvnTest="mardia" multivariatePlot="qq"

1.9 R for EDA : Practice Time

[R-code 1.8.1] iris-chisqplot.R

```
data(iris)
     setosa = iris[1:50, 1:4]   # Iris data only for setosa
    #versicolor = iris[51:100, 1:4] # Iris data only for versicolor
    #virginica = iris[101:150, 1:4] # Iris data only for virginica

# Chi-squre Plot for Checking MVN
     x=setosa
     n=dim(x)[[1]]
     p=dim(x)[[2]]
     S=cov(x)
     xbar=colMeans(x)
     m=mahalanobis(x, xbar, S)
     m=sort(m)
     id=seq(1, n)
     pt=(id-0.5)/n
     q=qchisq(pt, p)
     plot(q, m, pch="*", xlab="Quantile", ylab="Ordered Squared Distance")
     abline(0, 1)
# Correlation Coefficient Test for Normailty
    rq=cor(cbind(q, m))[1,2]
    rq
```

1.9 R for EDA : Practice Time

[R-code 1.8.2] iris-MVNtest.R

```
library("MVN")
iris
# MVN tests based on the Skewness and Kurtosis Ststistics
par(mfrow=c(1, 3))
setosa=iris[1:50, 1:4] # Iris data only for setosa and four variables
versicolor=iris[51:100, 1:4] # Iris data only for versicolor and four variables
virginica=iris[101:150, 1:4] # Iris data only for virginica and four variables

result_setosa=mardiaTest(setosa, qqplot=TRUE)
result_versicolor=mardiaTest(versicolor, qqplot=TRUE)
result_virginica=mardiaTest(virginica, qqplot=TRUE)

result_setosa
result_versicolor
result_virginica
```

mvn(setosa, mvnTest="mardia", multivariatePlot="qq")