

Multivariate Normal Distribution

▪ Bivariate Normal

$$P(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{z}{2(1-\rho^2)}\right]$$

$$z \equiv \frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}$$

- marginal pdf of $x_i \sim \text{Normal}(\mu_i, \sigma_i)$
- conditional pdf of $x_1|x_2 \sim \text{Normal}$ (complicate)
- how to generate a bivariate normal
 - In HW#2, using two normal distribution

▪ Multivariate normal $X \sim \text{MN}(\underline{\mu}, \Sigma)$

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(\sqrt{2\pi})^p |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma(\underline{x}-\underline{\mu})}$$

- \underline{x} : 변수 벡터, $\underline{\mu}$: 평균 벡터, Σ : 공분산 행렬, T: transpose

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_p \end{pmatrix}, \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_p \end{pmatrix}$$

- marginal of $x_i \sim \text{Normal}(\mu_i, \sigma_i)$

▪ Multivariate normal using R

▪ bivariate normal PDF

```
> library(mvtnorm)
> mu<-c(6,5)
> sigma<-matrix(c(9,5,5,4),2,2)
> x<-seq(-3,15,by=0.1)
> y<-seq(-3,3,by=0.1)
> z<-matrix(NA,length(x),length(y))
> for(i in 1:length(x)){for(j in 1:length(y))
+ {z[i,j]<-dmvnorm(c(x[i],y[j]),mean=mu,sigma=sigma)}}
> z
```

	[,1]	[,2]	[,3]	[,4]
[1,]	1.342163e-05	1.708552e-05	2.157237e-05	2.701556e-05
[2,]	1.291883e-05	1.652038e-05	2.095385e-05	2.636053e-05

▪ generating bivariate normal

```
> sigma <- matrix(c(9,5,5,4),2,2)
> x <- rmvnorm(n=500, mean=c(6,5), sigma=sigma)
> x
```

	[,1]	[,2]
[1,]	2.28714753	2.06071117
[2,]	9.43585309	6.96827622

```
> names(x)
NULL
> x0<-data.frame(x)
> names(x0)
[1] "col1" "col2"
> fix(x0)
> names(x0)
[1] "y1" "y2"
```



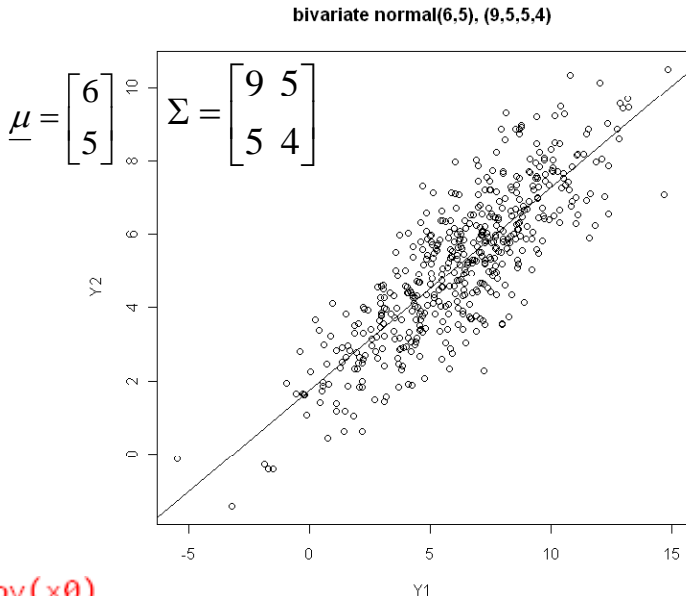
Bivariate normal Generating and scatter plot

scatter plot

```

> plot(x0$y1,x0$y2,main="bivariate normal(6,5), (9,5,5,4)",
+ xlab="Y1", ylab="Y2")
> abline(lm(x0$y2~x0$y1))

```



```

> cov(x0)
      y1      y2
y1 9.219004 5.260282
y2 5.260282 4.218760

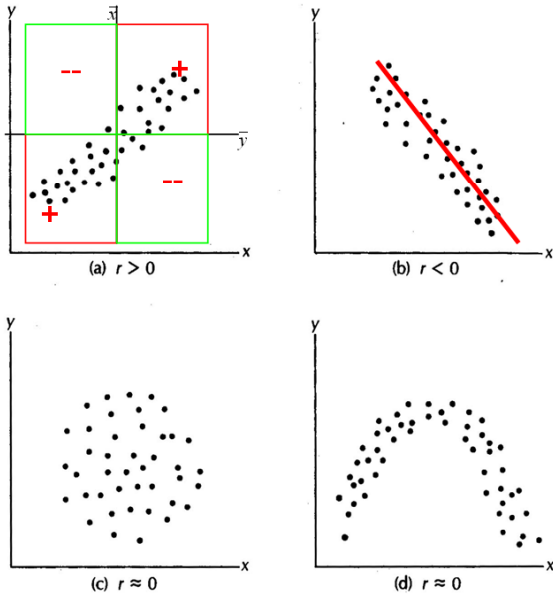
> cor(x0)
      y1      y2
y1 1.0000000 0.8328244
y2 0.8328244 1.0000000

> mean(x0)
      y1      y2
y1 6.222482 5.174247

```

산점도 scatter plot

- 각각의 점들은 짝으로 하여 2차원 공간에 표현
- 타원의 폭이 좁을수록 직선의 관계 높음
- 두 변수 간 함수 관계
- 개체 측면: 개체의 위치 표현, 개체의 그룹화 가능



상관계수

▪ Pearson 상관계수

- 두 변수의 선형(linear relationship)관계 정도, $-1 \leq r \leq 1$
- 타원의 폭이 짧고 길수록 상관계수 ± 1 에 가까움
- 직선의 경향, 점들이 직선에 가까울수록 상관 관계 높음

$$\begin{aligned}
 r &= \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \\
 &= \frac{E(X - E(X))E(Y - E(Y))}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \\
 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}
 \end{aligned}$$

▪ $H_0: \rho = 0$ $T = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t(n-2)$

▪ $H_0: \rho = \rho_0$ $T = 0.5 \ln \frac{1+r}{1-r} \sim N\left(0.5 \ln \frac{1+\rho_0}{1-\rho_0}, \frac{1}{n-3}\right)$

▪ 상관계수 해석

- 측정형(실험실) 데이터 0.7 이상
- 리커트 척도와 같은 순서형 데이터의 상관계수 매우 낮음
- 관측치가 많아지면 상관계수 높아짐
- 단순회귀의 결정계수(Determinant Coeff.) 제공근

▪ Non-parametric Correlation

- 관측치의 개수가 10~15개 미만이거나 관측치가 가질 수 있는 값의 수준이 5~10개 미만

Spearman 순위

$$\begin{aligned}
 r_s &= \frac{\sum (R_x - \bar{R}_x)(R_y - \bar{R}_y)}{\sqrt{\sum (R_x - \bar{R}_x)^2} \sqrt{\sum (R_y - \bar{R}_y)^2}} \\
 &\approx 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
 \end{aligned}$$

R 은 관측치 순위, $d_i = R_{x_i} - R_{y_i}$ 이다.

Kendall의 τ

$$\tau = \frac{\sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)}{\sqrt{(T_0 - T_x)(T_0 - T_y)}}$$

$$\text{sign}(w) = \begin{cases} -1, & w < 0 \\ 0, & w = 0 \\ 1, & w > 0 \end{cases} \quad T_0 = n(n-1)/2, \quad T_x = \sum t_i(t_i-1)/2,$$

t_i 는 동일한 x_i 의 i -번째 그룹 내의 관측치 개수이다.



상관계수 활용 예제 데이터

- 상관계수는 두 변수간의 선형 관계 정도 표현
- 상관계수가 높은 변수는 유사 개념을 측정한다. 즉 유사한 변수로 묶을 수 있다. (Spearman의 생각)

예제 데이터 Applicant.txt

- 지원자 48명 중 우수 지원자 5명을 선발하고자 15개 항목에 대해 평가.
- 측정 항목이 2개? 산점도 활용

```
> app<-read.table("APPLICANT.TXT",header=T)
> fix(app)
```

R Data Editor						
	ID	X1.FL.	X1.APP.	X3.AA.	X4.LA.	X
1	1	6	7	2	5	8

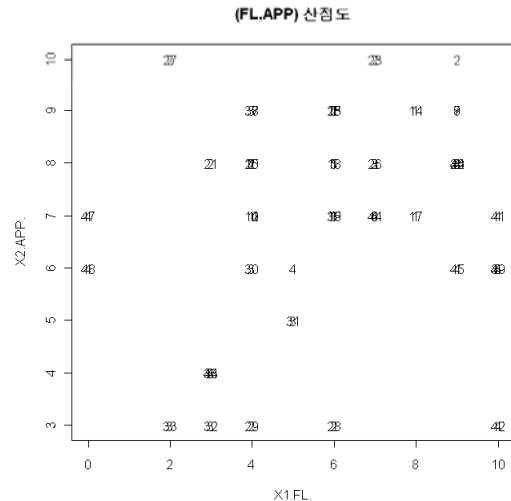
```
> attach(app)
> plot(X1.FL.,X2.APP.,pch=ID)
> text(X1.FL.,X2.APP.,ID)
```

In R

```
> cor.test(x0$y1,x0$y2)
```

Pearson's product-moment correlation

```
data: x0$y1 and x0$y2
t = 33.5754, df = 498, p-value < 2.2e-16
alternative hypothesis: true correlation is n
95 percent confidence interval:
 0.8038389 0.8578644
sample estimates:
cor
0.8328244
```



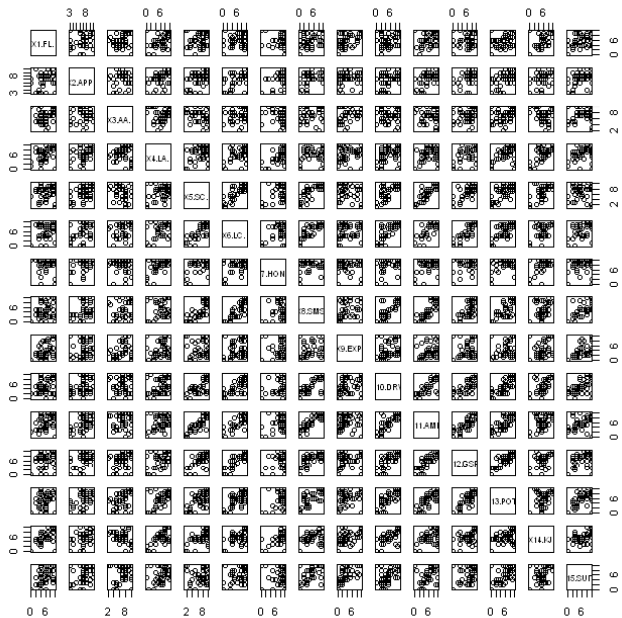
http://wolfpack.hnu.ac.kr



Scatter PLOT matrix

In R

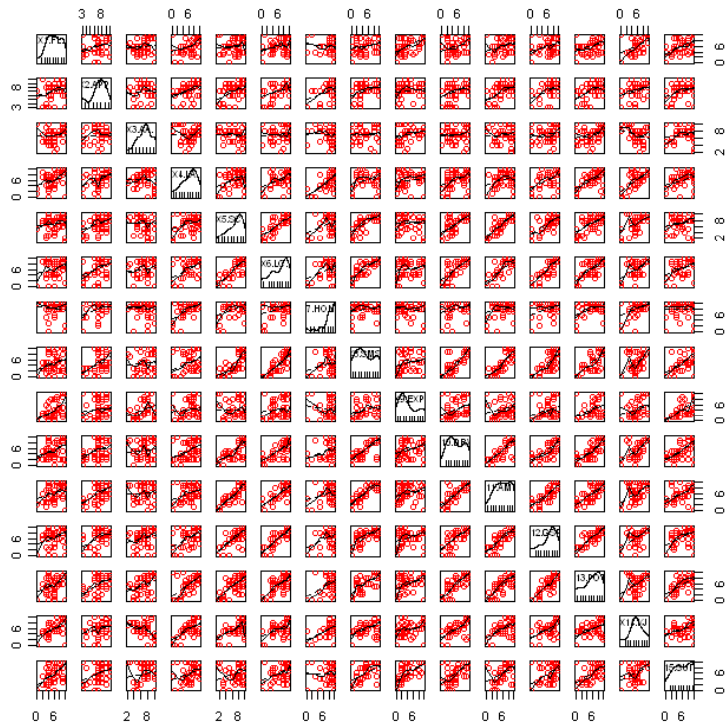
```
> app0 <- app[, 2:16]
> pairs(app0)
```



- 상관계수 만으로 변수를 그룹화 하는 것은 쉽지 않다.
 - 그룹 내 변수간 상관계수는 크고
 - 그룹 간 변수들과는 상관계수가 낮게...
 - 그러나 일부 변수들이 배신을 한다.

Another

```
> library(car)
> scatterplot.matrix(app0, labels=colnames(app0))
```



Correlation of Second order function data

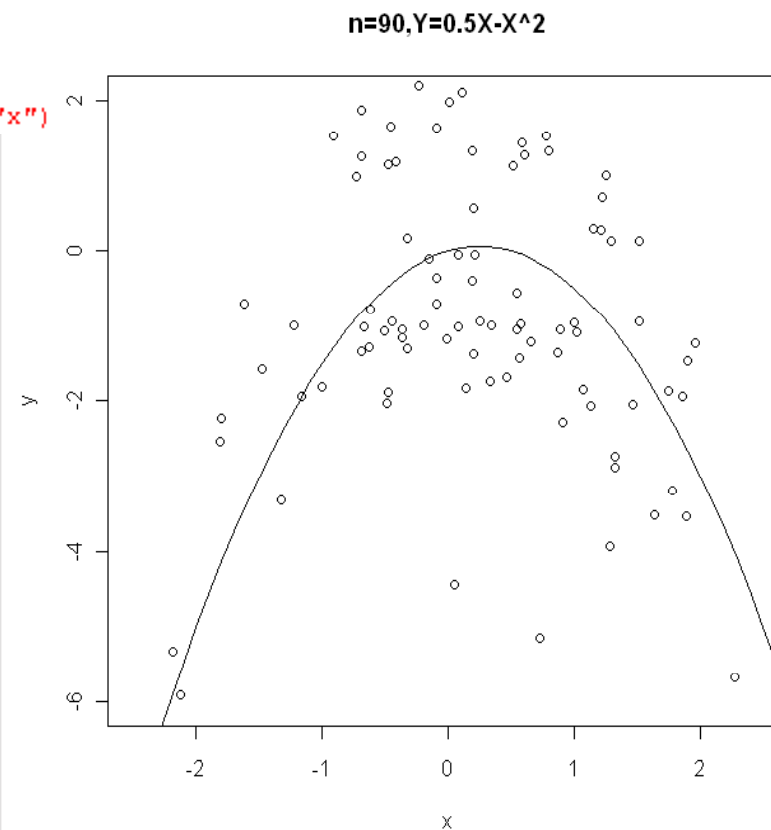
▪ R 프로그램

```
> split.screen(c(1,1))
> screen(1)
> plot(x, y, type='l', ylim=c(-6,2), xlim=c(-2.5,2.5),
+ ylab="y", xlab="x", main="n=90, Y=0.5X-X^2")
> screen(1)
> plot(xy, ylim=c(-6,2), xlim=c(-2.5,2.5), ylab="y", xlab="x")
```

▪ split.screen 스크린 분할

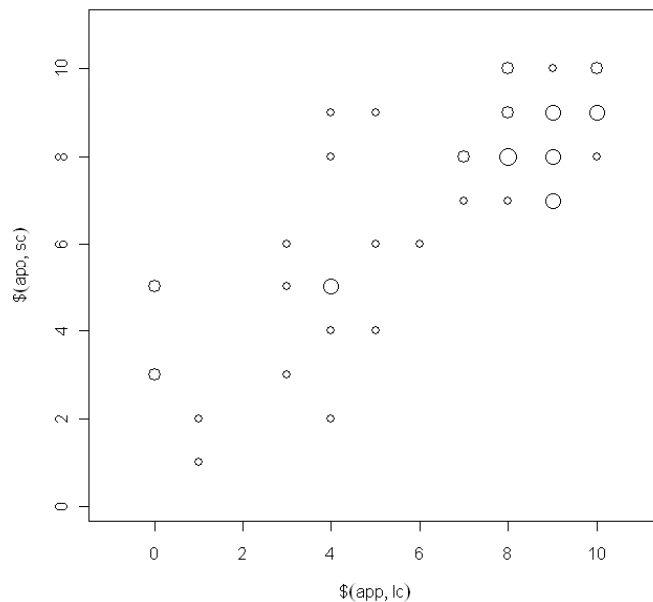
- c(행,열)
- screen(1)
 - 제일 스크린에 그래프 그리기

▪ scatter plot



Bubble plot in R

- X-축: LC, y-축: SC
- bubble의 크기: SMS
- 데이터 app에는 LC, SC, SMS 3개 변수만 있어야 한다.



```
> app<-read.table("app2.csv", header=T, sep=",")
> bubble.plot<-function(x, y, scale=0.1, xlab=substitute(x), ylab=substitute(y), ...){ z<-table(x, y)
+ xx<-rep(as.numeric(rownames(z)), ncol(z))
+ yy<-sort(rep(as.numeric(colnames(z)), nrow(z)))
+ id<-which(z!=0)
+ symbols(xx[id], yy[id], inches=F, circles=sqrt(z[id])*scale, xlab=xlab, ylab=ylab, ...)}
> bubble.plot(app$lc, app$sc)
```

HW#3 due 10.27.2008(Mon)

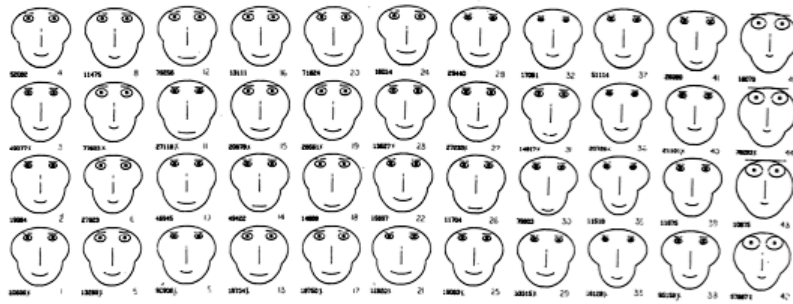
- ①MDA Term project: 서론 (2 페이지) 및 데이터 행렬
- ②페이지 67 #2
- ③페이지 67 #3



Chernoff face using R

정의

- Chernoff faces display multivariate data in the shape of a human face.. Each unit is represented as a schematic face. Variables of interest are represented by particular parameters of the face, e.g. the nose size, eye-to-eye distance, etc.



Using R

