Multivariate Normality
&
Quantile-Quantile Plot

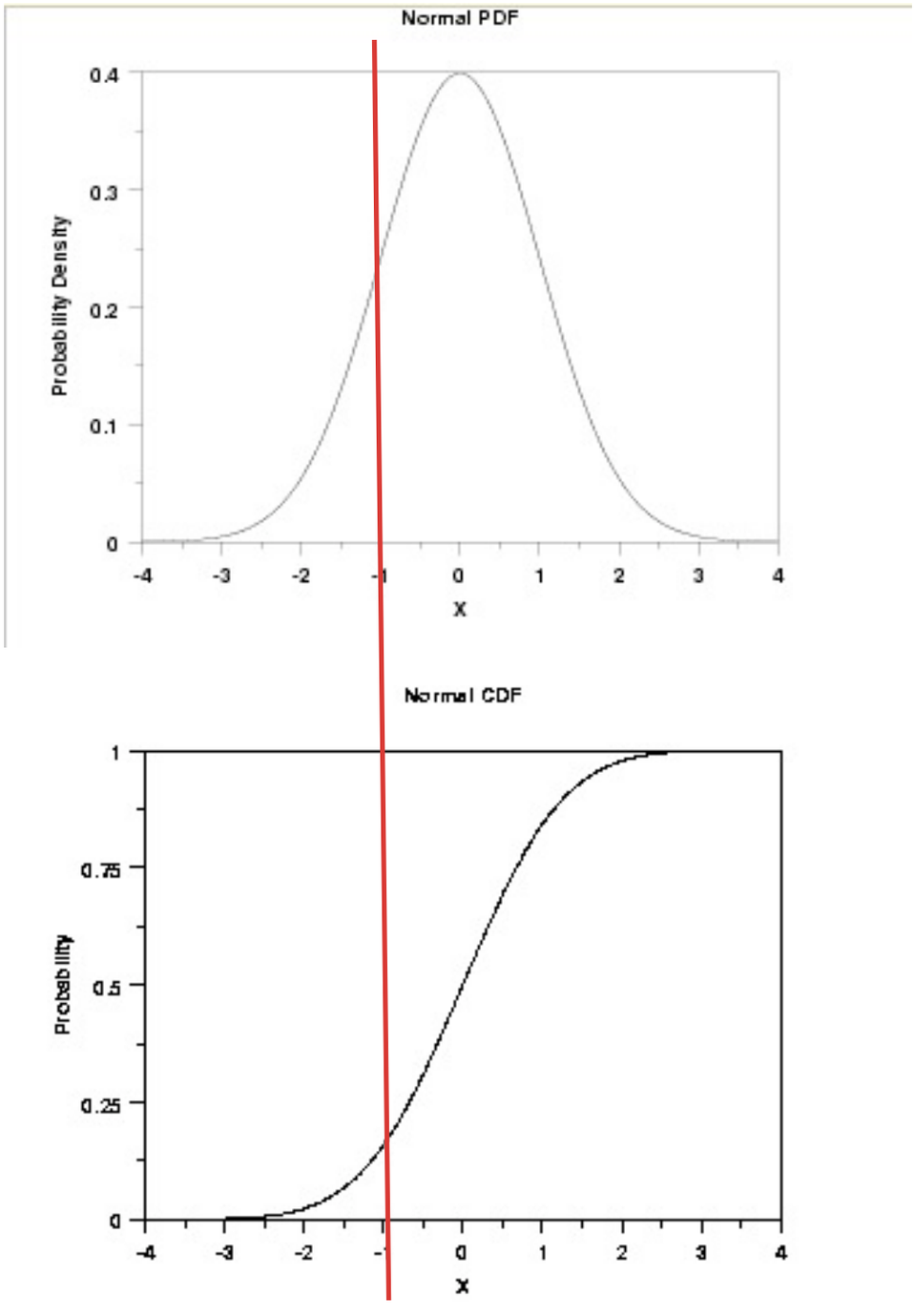
H. Park

HUFS

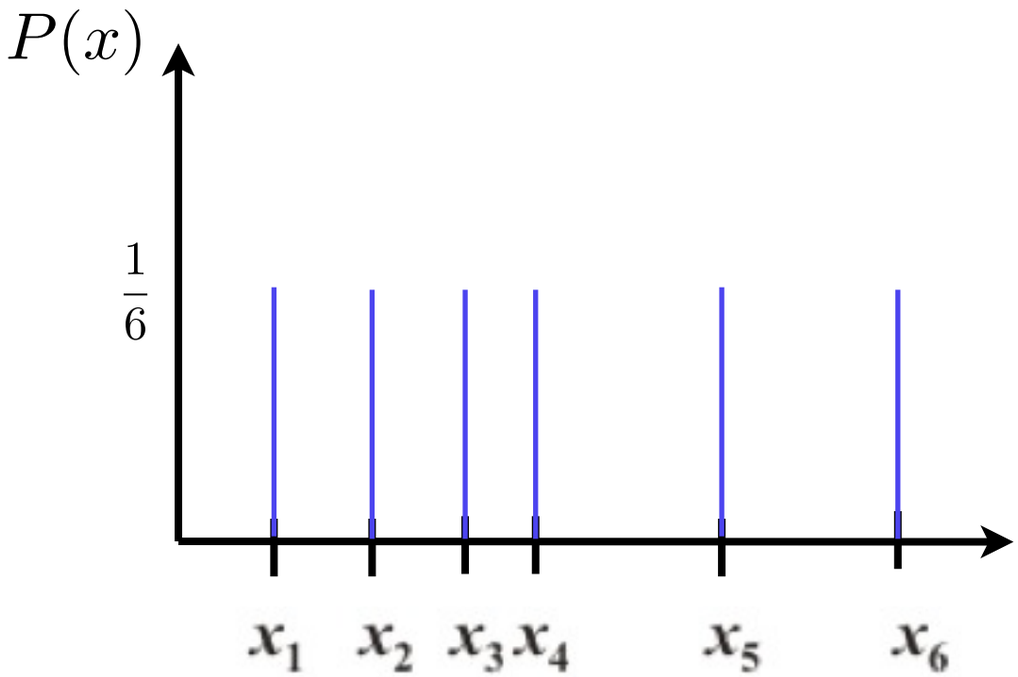
정규성 (Normality)

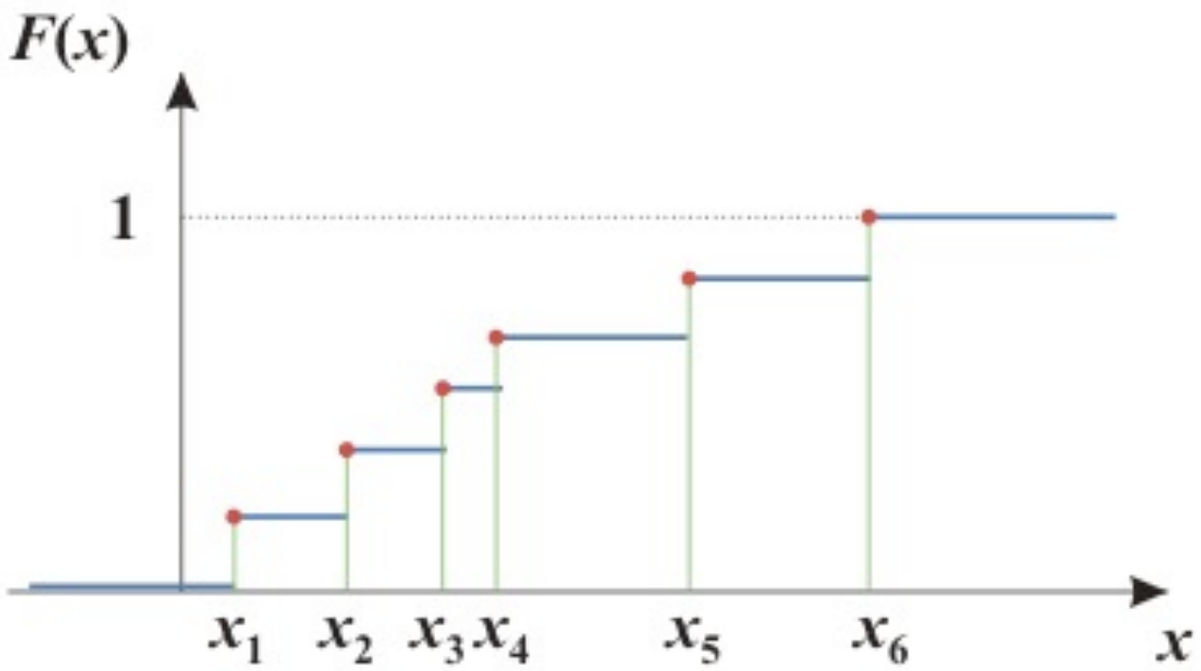
- 누적분포함수 (Cumulative DistributionFunction: c.d.f.)

$$f(x) = \text{p.d.f of x}$$
$$F(x) =$$



- 경험적 누적분포함수 (empirical c.d.f.)



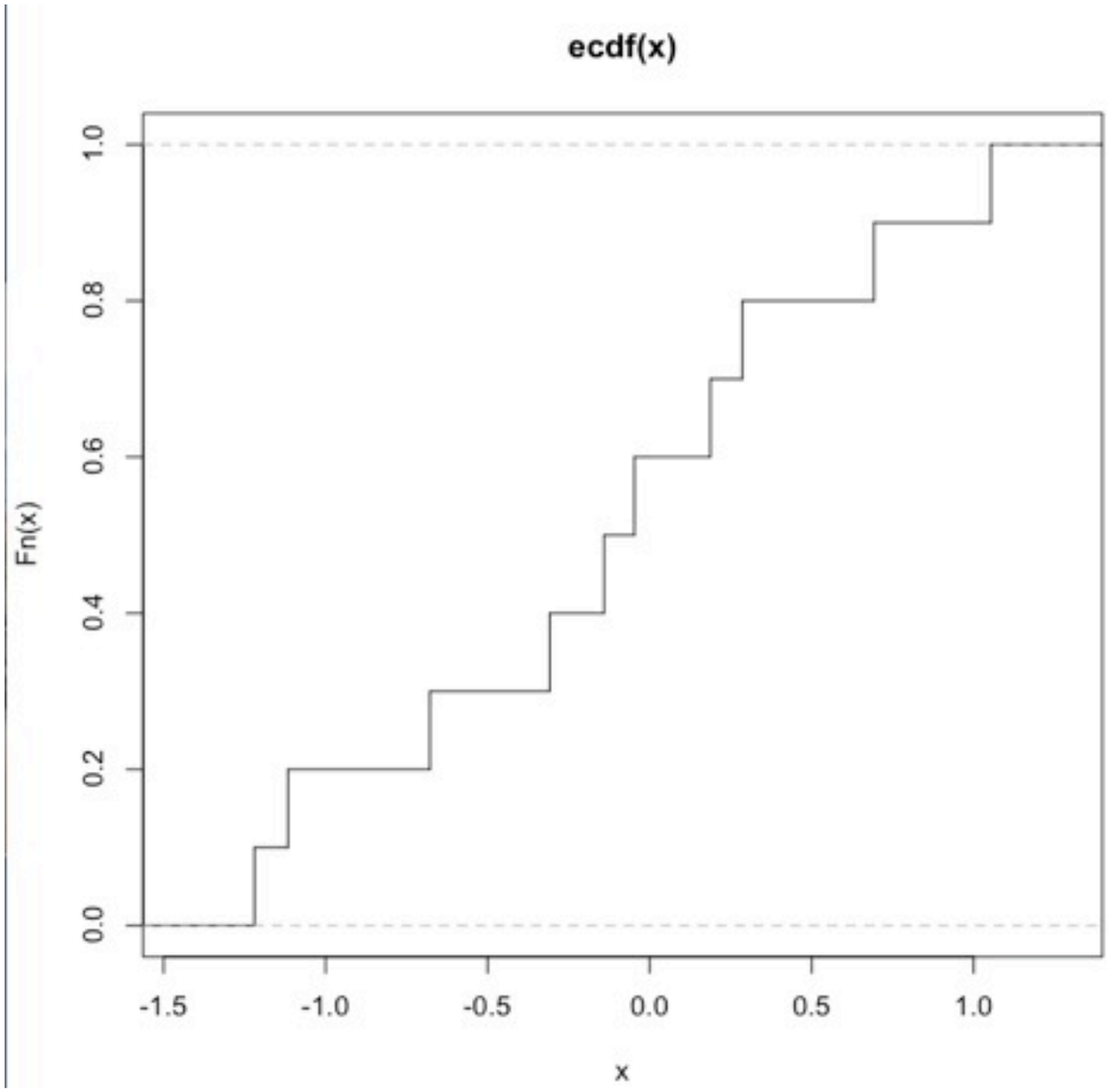


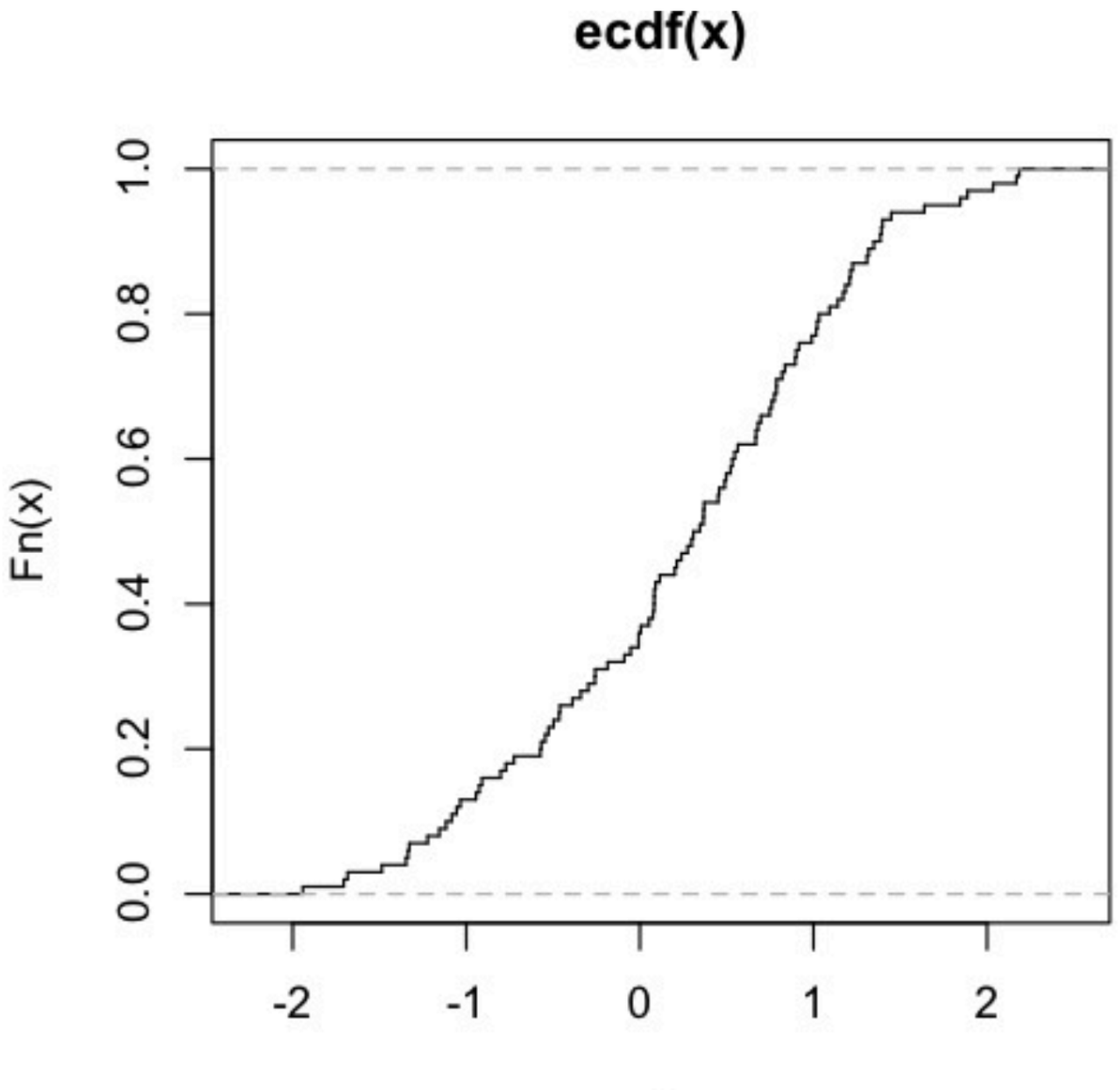
$$
\begin{aligned}
P(X < x_1) &= 0 \\
P(X \le x_1) &= \\
P(X < x_2) &= \\
P(X \le x_2) &= \\
P(X < x_3) &= \\
P(X \le x_3) &= \\
P(X < x_4) &= \\
P(X \le x_4) &= \\
P(X < x_5) &= \\
P(X \le x_5) &= \\
P(X < x_6) &= \\
P(X \le x_6) &=
\end{aligned}
$$

- 정규난수의 경험적 누적분포함수 (e.c.d.f.)

```
x<-rnorm(10)

f10<-ecdf(x)

plot(f10, x, verticals=T , do.p=F )
```

```
x<-rnorm(100)

f10<-ecdf(x)

plot(f10, x, verticals=T , do.p=F )
```





Quantile-Quantile Plot (Q-Q plot)

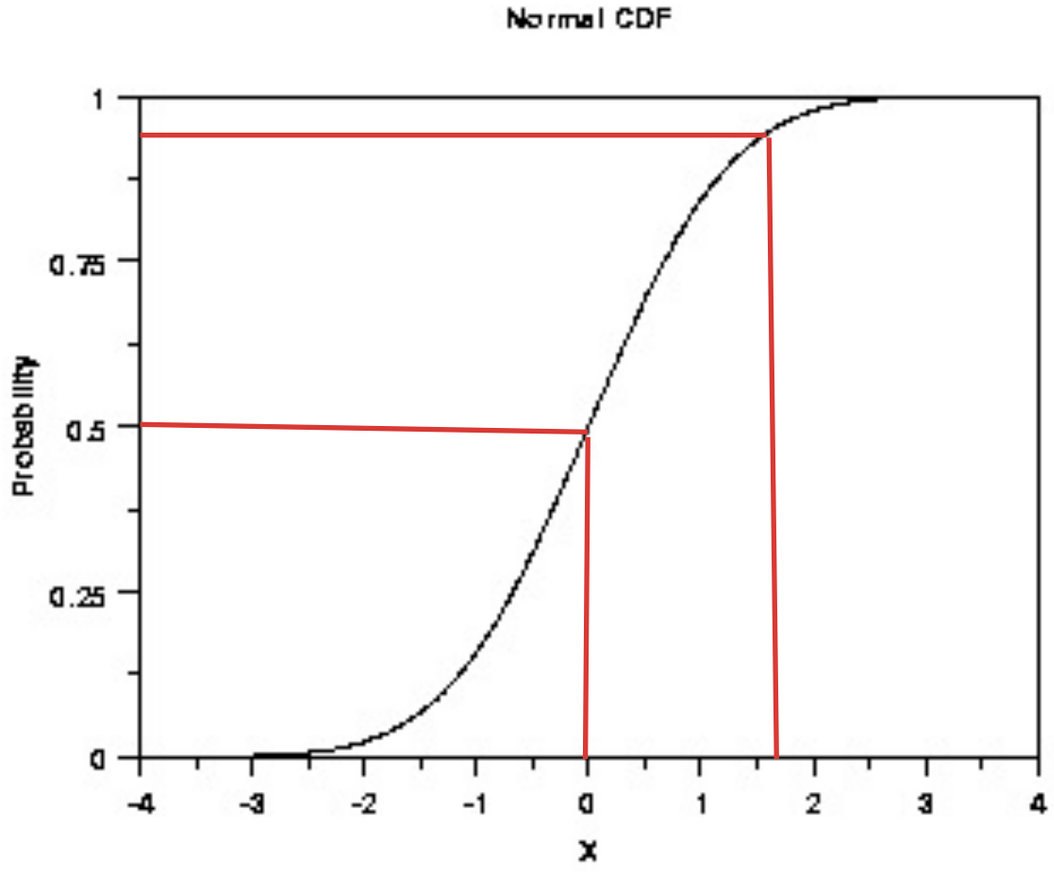
- 분위수(Quantile Point) 란?

→

- 표준정규분포의 분위수

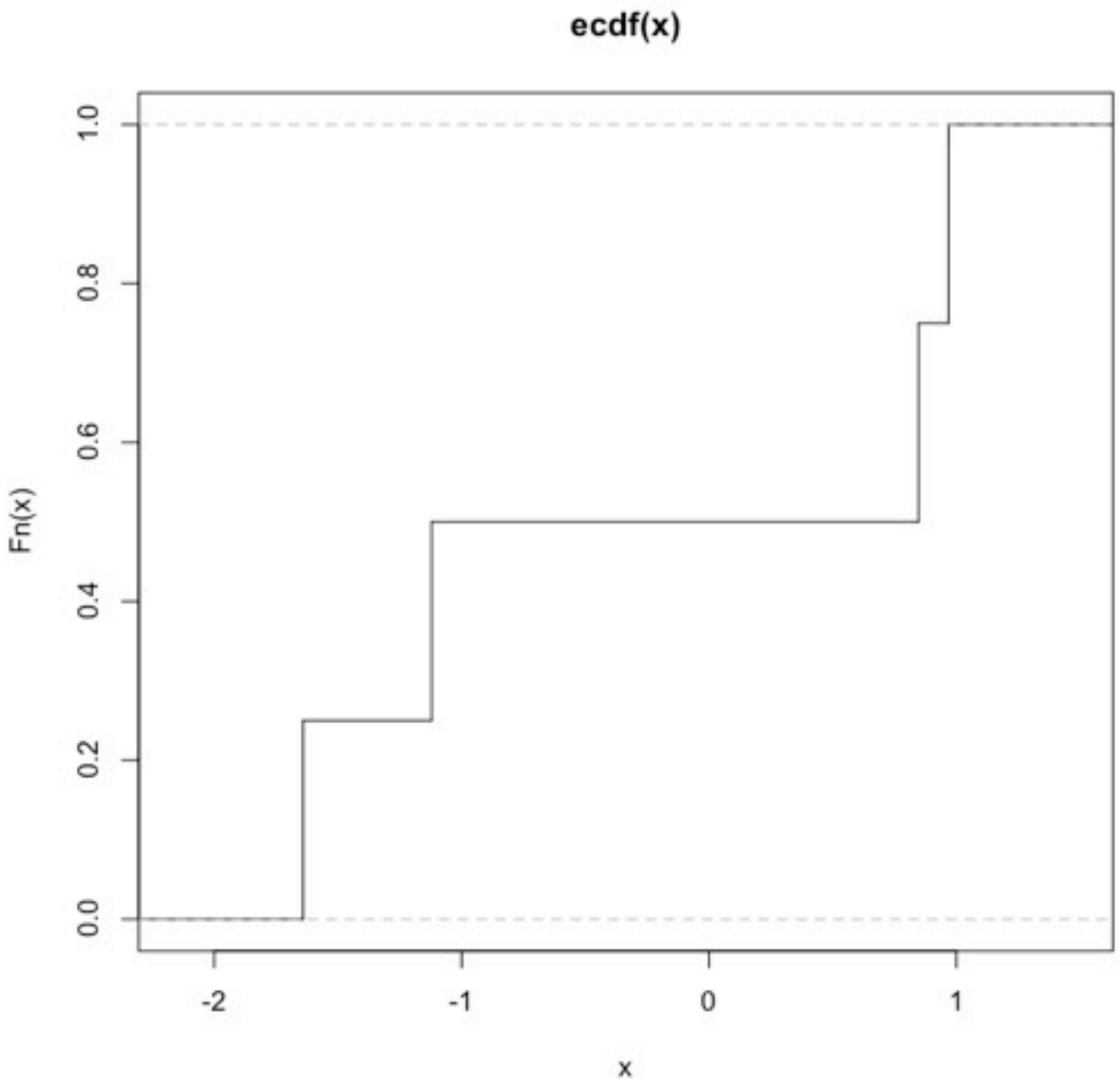
$$0 \quad = \quad F^{-1}(0.5)$$

$$1.64 \quad = \quad F^{-1}(0.95)$$

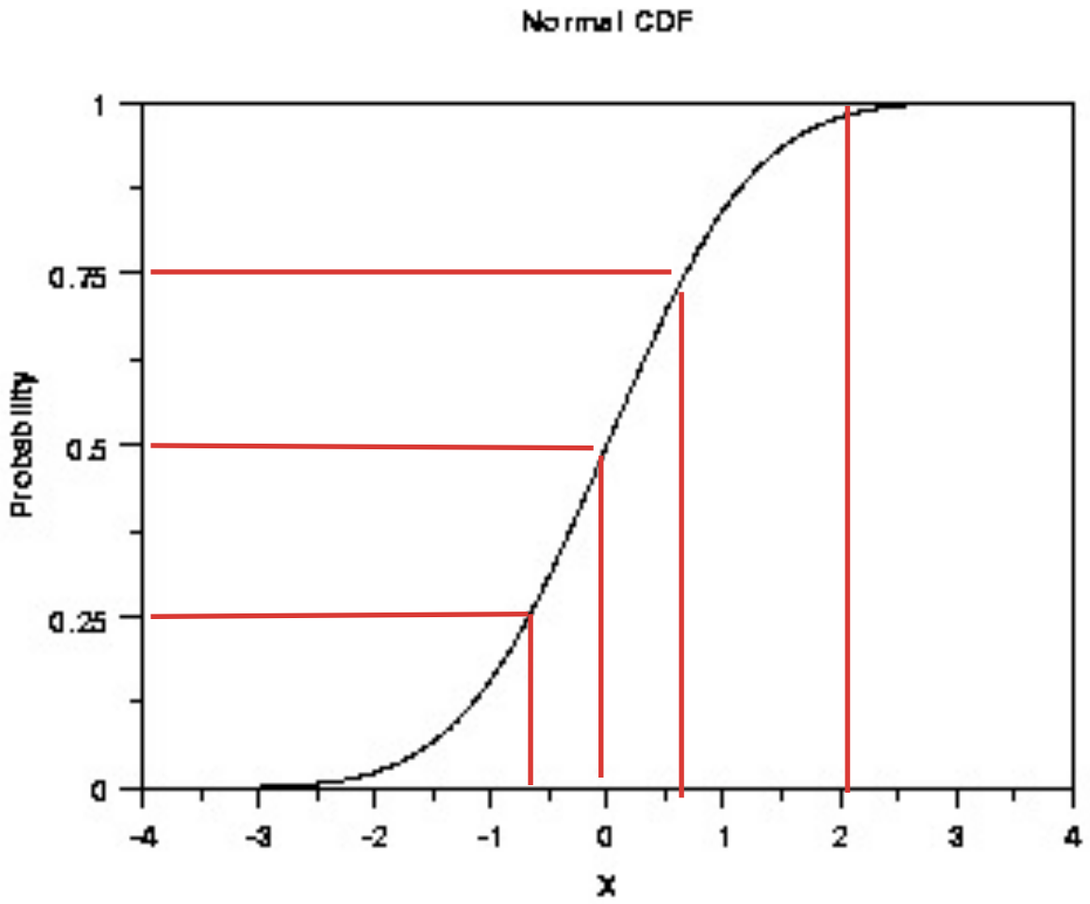


- (예) 아래 자료의 경험적 분위수를 구하라

$$x = -1.6,\ \ -1.1,\ 0.8,\ 0.9$$



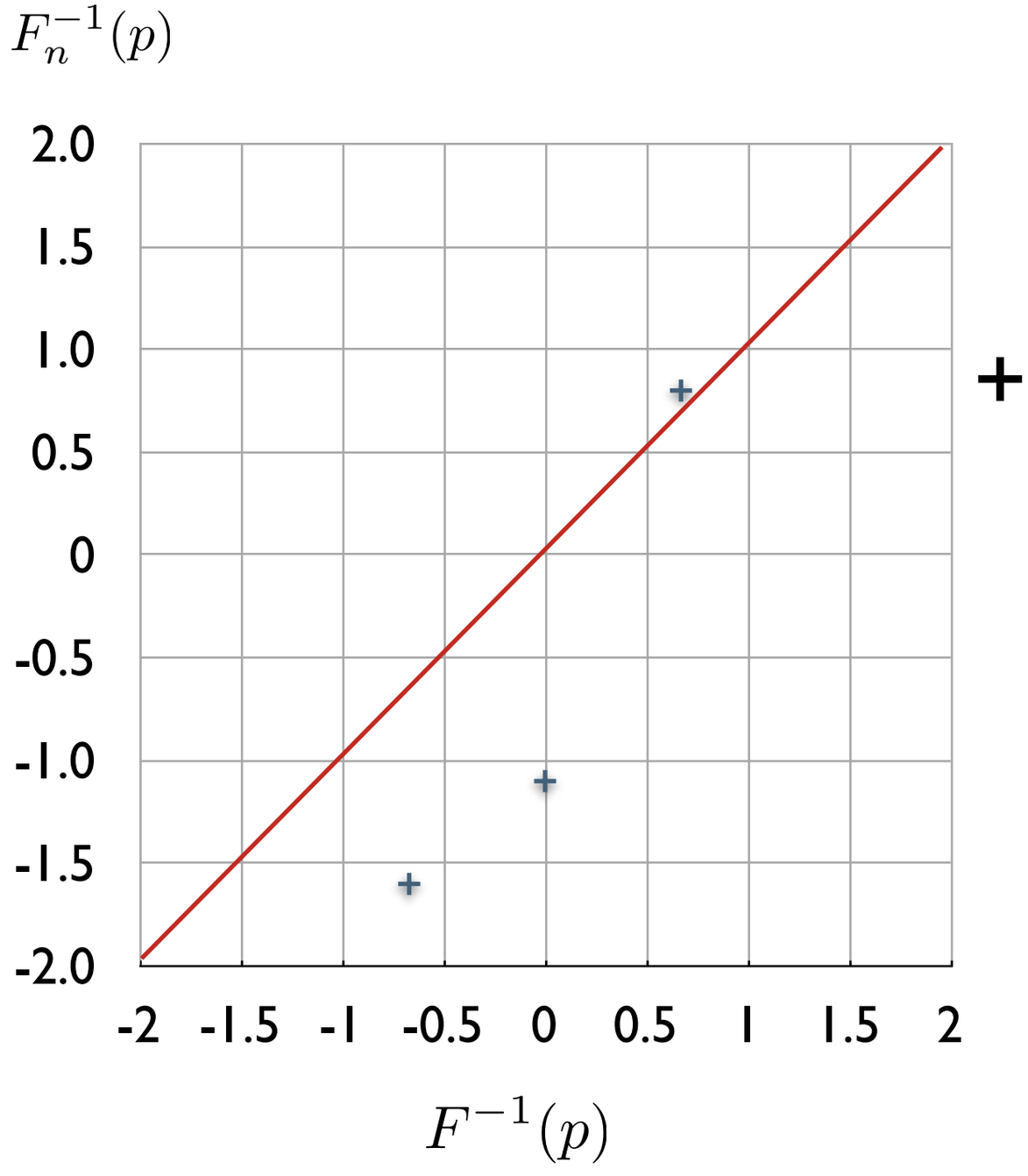
같은 자료를 가지고 표준정규분포에 의한
이론적 분위수를 구하라



0.25 분위수는
0.50 분위수는
0.75 분위수는
0.999 분위수는

경험적 분위수 vs. 이론적 분위수 (Q-Q plot)

	경험적	이론적
0.25 분위		
0.50 분위		
0.75 분위		
0.999 분위		



- 45도 =>

Q-Q plot

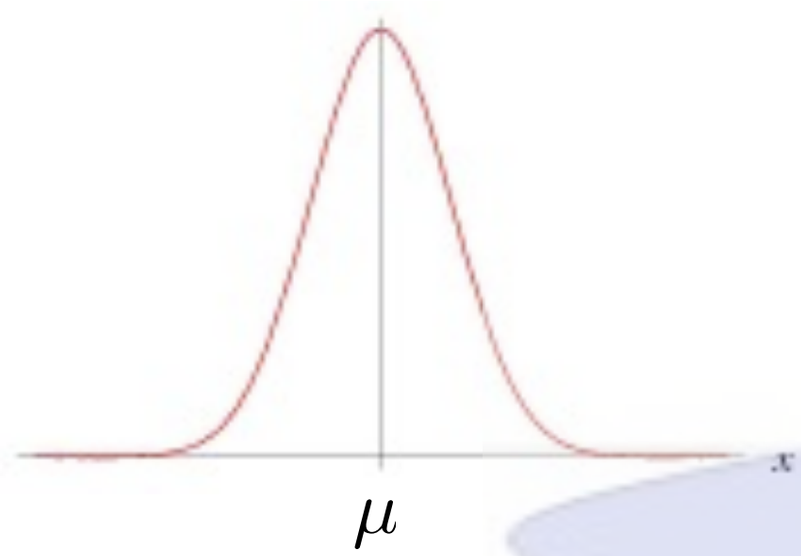
```
qqnorm(x)

abline(x)
```

```
qqplot(rnorm(1000, mean=0, sd=1), x)

abline(0,1)
```

qqline(x) 

$$x \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \Sigma)$$

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

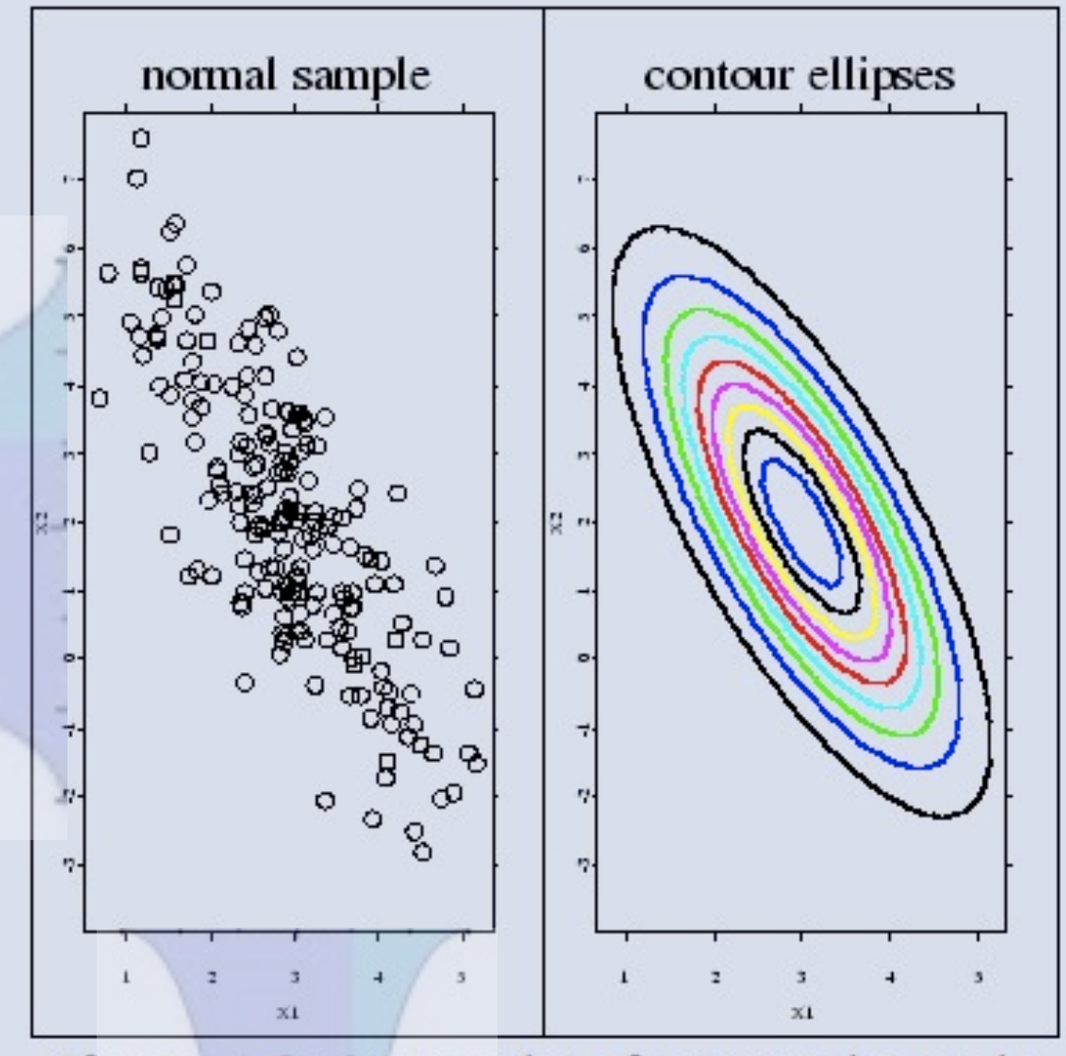


Figure 4.3: Scatterplot of a normal sample and contour ellipses for $\mu = \binom{3}{2}$ and

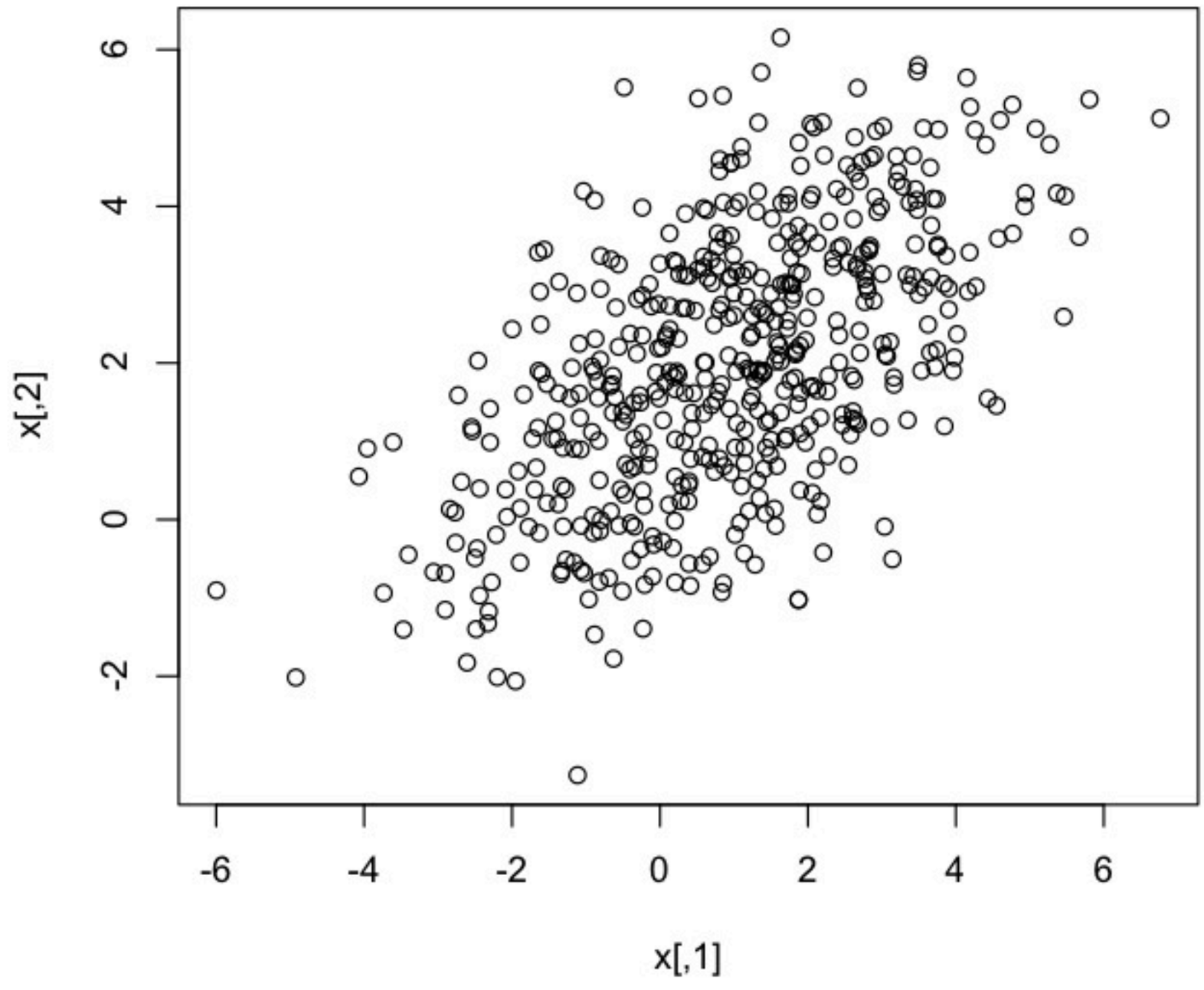
$$\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}.$$ MVAcontnorm.xpl

p=2

bivariate normal random number

```
sigma <- matrix(c(4,2,2,3), ncol=2)
x <- rmvnorm(n=500, mean=c(1,2), sigma=sigma)
colMeans(x)
var(x)
plot(x)
```

"mvtnorm 패키지 설치"



다변량 정규분포 점검

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{pmatrix} \qquad \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iq} \end{pmatrix} \qquad \bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_q \end{pmatrix}$$

If $\boldsymbol{x}_i \sim N(\boldsymbol{\mu}, \Sigma)$

- Empirical Quantile from d_i^2 , i=1, 2, ... , n
- Theoretical Quantile from χ^2_q
- Mahalanobis distance

Home Work

- crime data 가 다변량 정규분포를 따르는지 알아보기 위해 다음을 수행하여라.

1) 각 변수가 정규분포를 따르는지 Q-Q plot 을 각각 수행해 답하라.

2) chi-square qqplot 을 수행하여 다변량 정규분포임을 점검하여라. (해당 R 코드를 적어라)