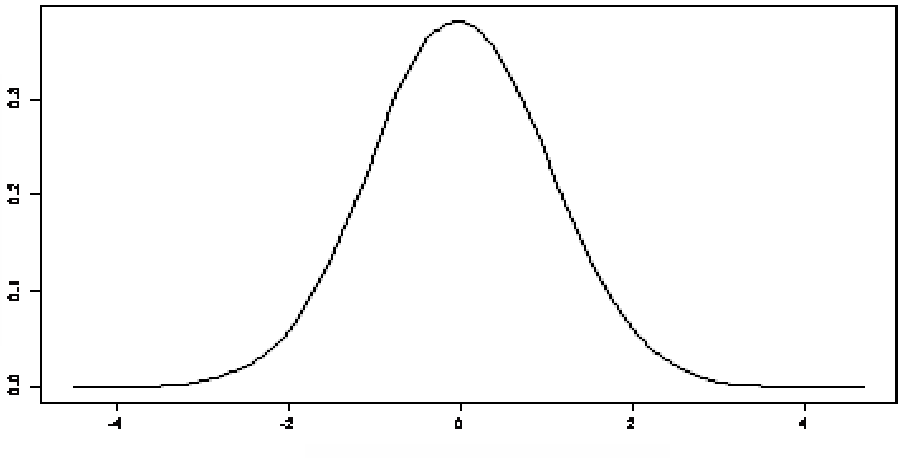
4장 다변량 정규분포

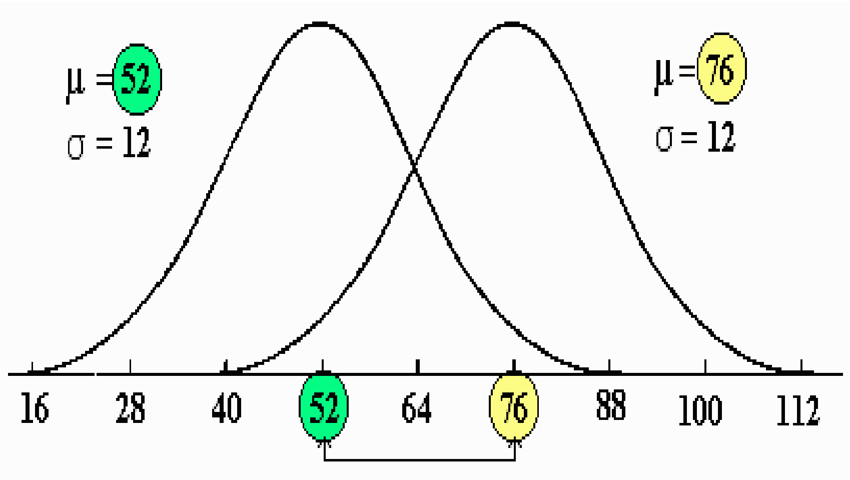
덕성여자대학교 정보통계학과 김 재희

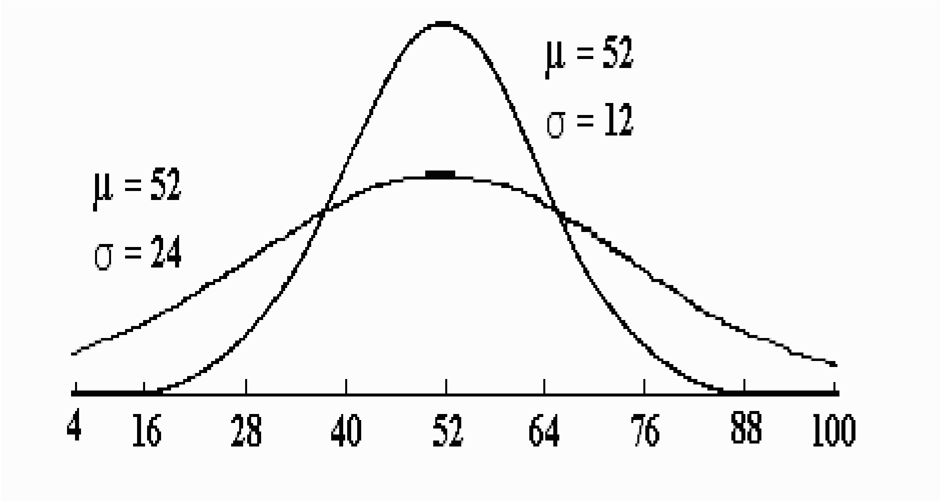
4.1 다변량 정규분포

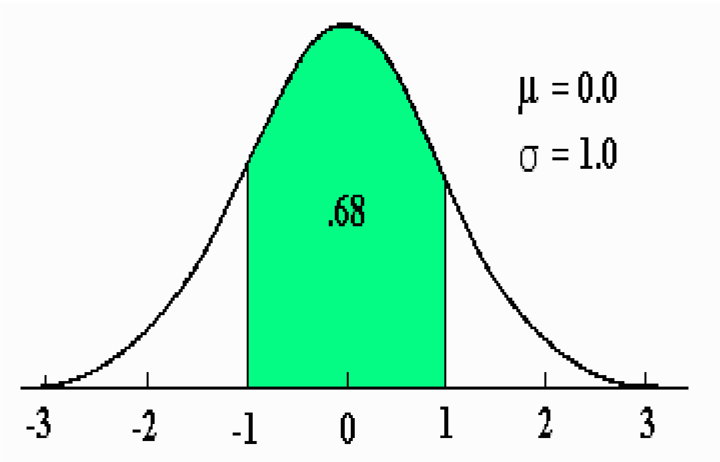
■ 평균 μ, 분산 σ^2인 정규분포를 따르는 일변량 확률변수 $X \sim N(\mu, \sigma^2)$ 확률밀도함수

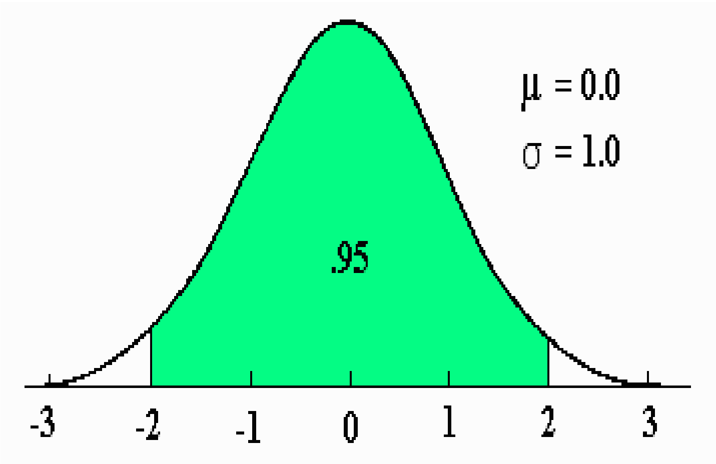
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty.$$




μ = 52
σ = 12
μ = 76
σ = 12
16
28
40
52
64
76
88
100
112


μ = 52
σ = 12
μ = 52
σ = 24
4
16
28
40
52
64
76
88
100


μ = 0.0
σ = 1.0
.68
-3
-2
-1
0
1
2
3


μ = 0.0
σ = 1.0
.95
-3
-2
-1
0
1
2
3

■ 지수함수 내에 있는

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

이 부분에 대해 p-차원으로의 확장

$$(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

가 일정한 값을 갖는다. 여기서 $\boldsymbol{\mu}= E(\boldsymbol{X})$, $\boldsymbol{\Sigma}= Cov(\boldsymbol{X})$

■ 일변량 정규 확률밀도함수에서의 상수항 $(2\pi)^{-1/2}(\sigma^2)^{-1/2}$는 p-차원의 체적(volume) 개념으로$(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}$이 얻어진다.

■ 평균벡터 $\boldsymbol{\mu}$, 공분산행렬 $\boldsymbol{\Sigma}$인 p-변량 정규분포를 따르는 확률변수 $\boldsymbol{X}$는 다음의 확률밀도함수를 가지며 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$로 표기한다:

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

$$\boldsymbol{x} = (x_1, \ldots, x_p)', \quad -\infty < x_i < \infty, \quad i = 1, 2, \cdots, p.$$

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

≪예제 4.1≫ 이변량 정규분포 확률밀도함수(bivariate normal density)

이변량 정규분포를 따르는 $\boldsymbol{X}=\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$가 평균벡터 $\boldsymbol{\mu}=\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$,

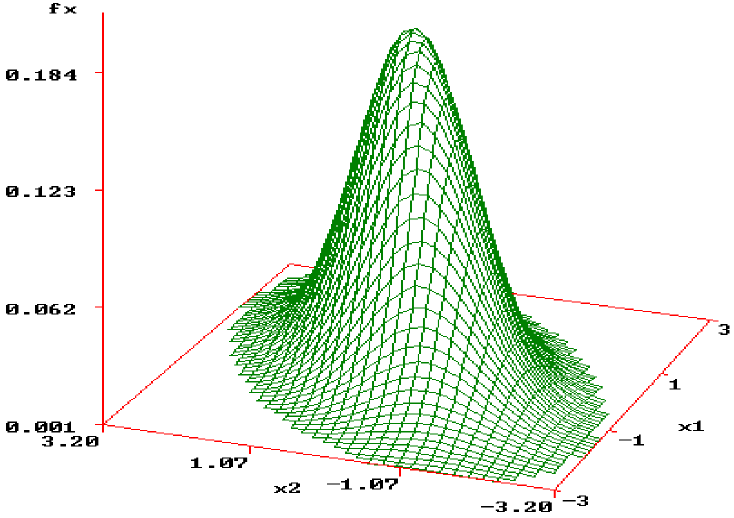
공분산행렬 $\boldsymbol{\Sigma}=\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$를 가질 때 확률밀도함수를 구해보자.

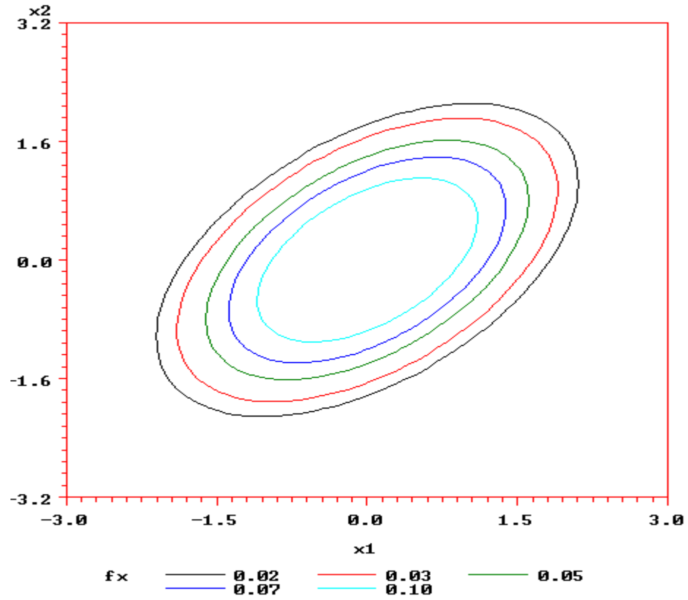
$$f(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho_{12}^2)}}$$

$$\times \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{\mathrm{x}_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2+\left(\frac{\mathrm{x}_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2-2\rho_{12}\left(\frac{\mathrm{x}_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{\mathrm{x}_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\}$$

$$-\infty < x_1 < \infty, \quad -\infty < x_2 < \infty.$$

여기서 $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} = corr(X_1, X_2)$





[그림 4.1] 상관계수 $\rho = 0.5$를 갖는 이변량 정규분포의 확률밀도함수

[그림 4.2] 상관계수 $\rho = 0.5$인 이변량 정규분포의 확률밀도함수의 등고선그림

4.2 다변량 정규분포의 성질

$X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$일 때,

<u>성질 4.1</u> (선형결합식의 분포)

(i) 상수벡터 $\boldsymbol{a} = (a_1, a_{2,} \ldots, a_p)'$ 에 대해 선형결합식(linear combination) $\boldsymbol{a}'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$는 평균 $\boldsymbol{a}'\boldsymbol{\mu}$, 분산 $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$를 갖는 일변량 정규분포를 따른다.

$$\boldsymbol{a}'\boldsymbol{X} \sim N(\boldsymbol{a}'\boldsymbol{\mu},\ \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$$

<u>성질 4.2 (표준화 변수)</u>

표준화 벡터(standardized vector)

$$\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$$

는 $N_p(\mathbf{0}, \boldsymbol{I})$을 따른다. 여기서 $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$이며 $\boldsymbol{\Sigma}^{1/2}$는 $\boldsymbol{\Sigma}$의 제곱근행렬

성질 4.3 (주변 분포)

p-변량 정규분포를 따르는 확률벡터 $\boldsymbol{X}$를 $q\times 1$벡터 $\boldsymbol{X}_1$과 $(p-q)\times 1$벡터 $\boldsymbol{X}_2$로 분할할 때

$\boldsymbol{X}_{p\times 1}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}$는 평균벡터 $\boldsymbol{\mu}_{p\times 1}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}$와 공분산행렬 $\boldsymbol{\Sigma}_{p\times p}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$
을 가지며 $\boldsymbol{X}_1$은 q-변량 정규분포를 따른다. 즉 $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$이다.

성질 4.4 (조건부 분포)

$\boldsymbol{X}_2=\boldsymbol{x}_2$가 주어졌을 때 $\boldsymbol{\Sigma}_{22}\neq \mathbf{0}$일 때,

$$\boldsymbol{X}_1|\boldsymbol{X}_2=\boldsymbol{x}_2 \sim N(\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

성질 4.5 (정규 확률벡터 합의 분포)

$\boldsymbol{X}_{p\times 1}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}$가 평균벡터 $\boldsymbol{\mu}_{p\times 1}=\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}$와 공분산행렬 $\boldsymbol{\Sigma}_{p\times p}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{pmatrix}$을 가질 때

$$\boldsymbol{X}_1+\boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{11}+\boldsymbol{\Sigma}_{22})$$

<u>성질 4.6 (독립성)</u>

(i) 확률벡터 $\boldsymbol{X}$의 부분벡터인 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$에 대해,
공분산행렬 $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$이면 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$는 서로 독립이다.

<u>성질 4.7 (카이제곱 분포)</u>

$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$이며 $|\boldsymbol{\Sigma}| > 0$일 때

(i) $(\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_p^2$

(ii) $P\left\{\boldsymbol{X}: (\boldsymbol{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \leqq \chi_p^2(\alpha)\right\} = 1 - \alpha$

4.3 다변량 정규분포의 모수 추정

정리 4.10 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n \sim iid\ N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$일 때 $\boldsymbol{\mu},\boldsymbol{\Sigma}$의 최대우도추정량은

$$\hat{\mu}=\overline{\boldsymbol{X}}\ ,\quad \hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i-\overline{\boldsymbol{X}})(\boldsymbol{X}_i-\overline{\boldsymbol{X}})'=\frac{n-1}{n}\boldsymbol{S}$$

$\overline{\boldsymbol{X}},\ \hat{\Sigma}$: 충분통계량

정리 4.11 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n \sim iid\ N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$일 때

$\overline{\boldsymbol{X}}$ 와 $\boldsymbol{S}$ 의 표본분포(sampling distribution)

(i) $\overline{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu},\frac{1}{n}\boldsymbol{\Sigma}\right)$

(ii) $(n-1)\boldsymbol{S} \sim$ 자유도 $n-1$인 위샤트 분포(Wishart distribution)

(iii) $\overline{\boldsymbol{X}}$와 $(n-1)\boldsymbol{S}$는 서로 독립이다.

(iv) $(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\left(\frac{\boldsymbol{\Sigma}}{n}\right)^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}) \sim \chi_p^2$

(v) $n-p\rightarrow\infty$이면 $(\overline{\boldsymbol{X}}-\boldsymbol{\mu})'\left(\frac{\boldsymbol{S}}{n}\right)^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu})$는 근사적으로 χ_p^2를 따른다.

4.4 위샤트 분포

- $\boldsymbol{Z}_j \sim iid\ N_p(\mathbf{0},\boldsymbol{\Sigma})$일 때, 확률벡터의 곱인 $\sum_{j=1}^{n}\boldsymbol{Z}_j\boldsymbol{Z}_j{}'$는 자유도가 n인 위샤트분포(Wishart distribution)를 따름. $W_p(n,\boldsymbol{\Sigma})$로 표기.

- 자유도가 $(n-1)$인 위샤트 분포(Wishart distribution)를 따르는 확률행렬 $\boldsymbol{A}$의 확률밀도함수

$$W_{n-1}(\boldsymbol{A}|\boldsymbol{\Sigma}) = \frac{|\boldsymbol{A}|^{(n-p-2)/2}e^{-tr(\boldsymbol{A}\boldsymbol{\Sigma}^{-1})/2}}{2^{p(n-1)/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{(n-1)/2}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n-i)\right)}$$

여기서 $\boldsymbol{A}$는 양정치행렬, $\Gamma(\cdot)$는 감마함수.

■ Wishart 분포를 따르는 확률행렬의 분포에 관한 성질

(i) 상수행렬 C에 대해 CAC'는 $W_p(n-1, C\Sigma C')$를 따른다.

(ii) $\boldsymbol{A}_i \sim W_{p_i}(n_i, \boldsymbol{\Sigma})$일 때, 이들 확률행렬의 합 $\boldsymbol{A}$는 다음의 위샤트분포를 따른다.

$$\boldsymbol{A} = \sum_{i=1}^{q} \boldsymbol{A_i} \sim W_{\sum_{i=1}^{q} n_i} (\boldsymbol{A} | \boldsymbol{\Sigma})$$

정리 4.12 (대표본 분포) $p \times 1$ 확률벡터로 구성된 확률표본 $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$가 서로 독립이고

평균벡터 $\boldsymbol{\mu}$, 공분산행렬 $\boldsymbol{\Sigma}$인 동일한 다변량 분포를 따를 때

(i) n이 충분히 크고 n이 p보다 훨씬 크면, $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.

(ii) $n-p$가 충분히 크면, $n(\overline{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\overline{\boldsymbol{X}} - \boldsymbol{\mu})$는 자유도가 p인 카이제곱 분포를 따름.

4.5 정규성 평가

관측벡터 $\boldsymbol{X}$ 가 p-변량 정규분포를 따른다는 가정을 검토하고자 할 때

(1) $\boldsymbol{X}$의 주변분포가 모두 정규분포인가

(2) $\boldsymbol{X}$의 구성성분 $X_i,\ \ i=1,2,\ ...,p$들의 선형결합식의 분포가 모두 정규분포인가

(3) 두 개 구성성분들의 산점도(scatter plot)를 그려보면 타원형인가

4.5.1 일변량 변수에 대한 정규성 평가 : 분위수대분위수 그림(Q-Q plot)

4.5.2 다변량 변수벡터에 대한 정규성 평가: 카이제곱 그림

≪예제 4.2≫ A 지역의 주민을 대상으로 가장의 나이(age)와 수입(income)을 조사한 결과 얻은 [표 4.1] 자료에 대해 일변량 정규성을 점검하고자 한다.

[표 4.1] 나이와 연수입 자료

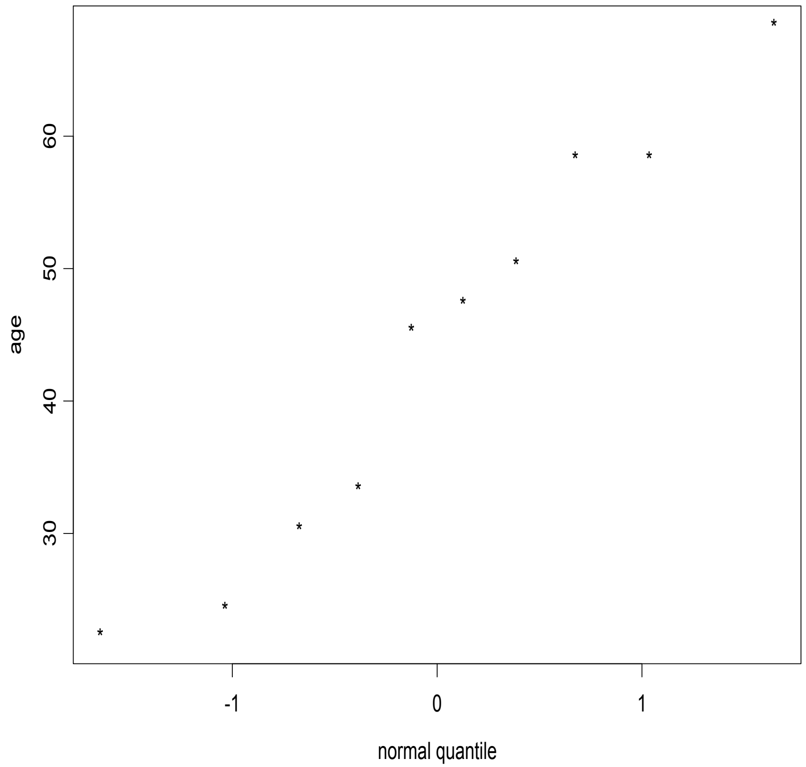
표본 번호	나이	수입 (단위:100만원)
1	68	4
2	58	45
3	45	35
4	50	25
5	33	28
6	24	11
7	58	10
8	22	9
9	47	20
10	30	35

$\boldsymbol{X}' = (\text{나이},\ \text{수입})$ 에 대한 표본평균벡터와 표본공분산행렬을 구하면 다음과 같다.

$$\overline{\boldsymbol{X}} = \begin{pmatrix} 43.5 \\ 22.2 \end{pmatrix},\ \boldsymbol{S} = \begin{pmatrix} 245.83 & 0.67 \\ 0.67 & 185.96 \end{pmatrix}$$

[표 4.2] 나이(age) 순서통계량과 정규분포 사분위수

순서 i	나이 관측값의 순서 통계량 $x_{(i)}$	경험 누적 확률 $\frac{i-0.5}{n}$	정규분포의 사분위수 $q_{(i)}$
1	22	0.05	-1.645
2	24	0.15	-1.036
3	30	0.25	-0.674
4	33	0.35	-0.385
5	45	0.45	-0.126
6	47	0.55	0.126
7	50	0.65	0.385
8	58	0.75	0.674
9	58	0.85	1.036
10	68	0.95	1.645



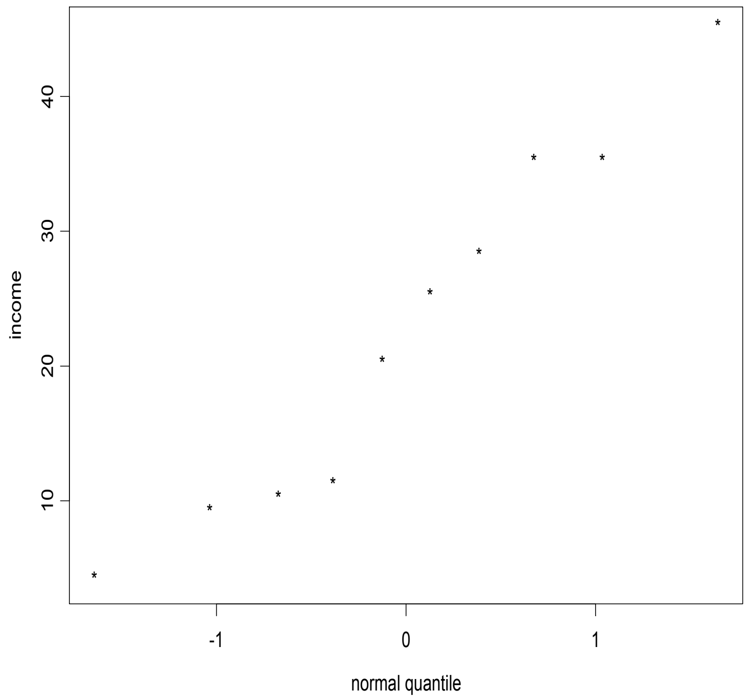
[그림 4.3] 나이에 대한 정규확률그림

[프로그램 4.1] Q-Q 그림

```
n=length(age)
t=seq(1,n)
age_s = sort(age)   ;  income_s = sort(income)
edf=(t -0.5)/n
q=qnorm( edf, 0, 1)
cbind( t, age_s, edf, q)      # 표 4.3
cbind( t, income_s, edf, q)   # 표 4.3

qqnorm(age)    # normal Q-Q plot    그림 4.3
qqline(age)    # normal quantile line

qqnorm(income)    # normal Q-Q plot  그림 4.4
qqline(income)    # normal quantile line
```

[표 4.3] 수입(income) 순서통계량과 정규분포 사분위수

순서 i	수입관측값의 순서통계량 $x_{(i)}$	경험 누적 확률 $\frac{i-0.5}{n}$	정규분포의 사분위수 $q_{(i)}$
1	4	0.05	-1.645
2	9	0.15	-1.036
3	10	0.25	-0.674
4	11	0.35	-0.385
5	22	0.45	-0.126
6	25	0.55	0.126
7	28	0.65	0.385
8	35	0.75	0.674
9	35	0.85	1.036
10	45	0.95	1.645



[그림 4.4] 수입에 대한 정규확률그림

4.5.2 다변량 변수벡터에 대한 정규성 평가

만약 관측벡터 $\boldsymbol{X}$가 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따른다면 다음의 거리(표본에 대한 마할라노비스 거리)

$$d^2 = (\boldsymbol{X} - \overline{\boldsymbol{X}})' \boldsymbol{S}^{-1} (\boldsymbol{X} - \overline{\boldsymbol{X}})$$

는 근사적으로 자유도가 p인 카이제곱분포를 따르게 된다.

카이제곱그림(chi-square plot)을 그리는 순서는 다음과 같다.

(1) d_i^2 을 구하고 순서를 매긴다.

$$d_i^2 = (\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}) , \quad i = 1, 2, \ldots, n,$$

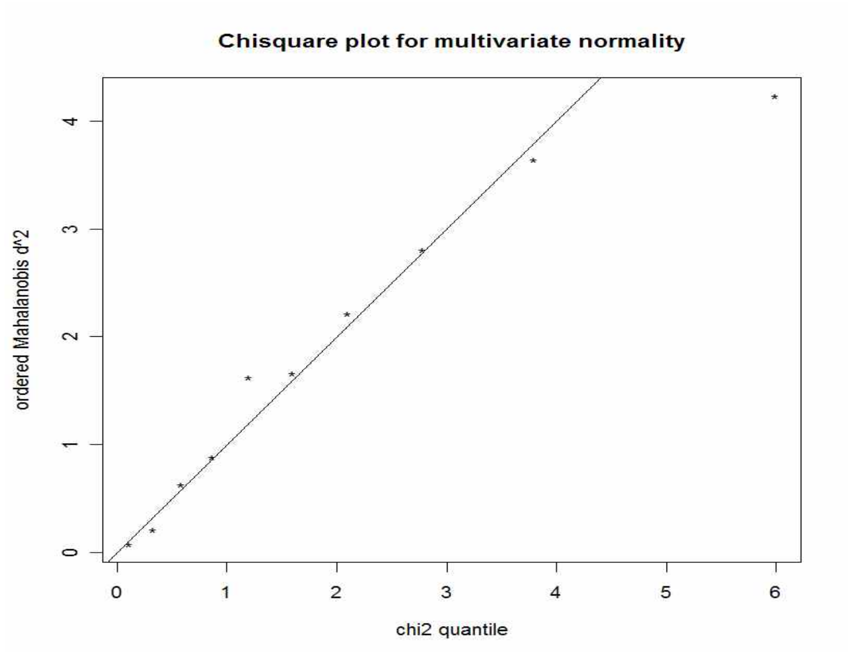
$$d_1^2 \leqq d_2^2 \leqq \cdots \leqq d_n^2$$

(2) $\left\{ \chi_p^2 \left(\frac{i - 0.5}{n} \right), \ d_i^2 \right\}$ 점을 그린다. 즉 d_i^2과 이론적으로 카이제곱분포를 따를 때의 사분위수를 구해 비교하여 다변량 정규성을 검토한다.

≪예제 4.4≫ (나이, 수입) 확률벡터가 이변량 정규분포를 따르는지 카이제곱그림을 그려보자.

[표 4.4] 마할라노비스 거리제곱 순서통계량과 카이제곱분포 사분위수

순서 i	관측벡터의 마할라노비스 거리 제곱에 대한 순서 통계량 d_i^2	경험 누적 확률 $\frac{i-0.5}{n}$	카이제곱분포의 사분위수 $q_{(i)}$
1	0.076	0.05	0.103
2	0.213	0.15	0.325
3	0.631	0.25	0.575
4	0.890	0.35	0.862
5	1.627	0.45	1.196
6	1.661	0.55	1.597
7	2.215	0.65	2.100
8	2.809	0.75	2.773
9	3.641	0.85	3.794
10	4.236	0.95	5.991



[그림 4.5] (나이, 수입)에 대한 카이제곱그림

```
  # chisquare plot     [프로그램 4.2] 카이제곱 그림
chi2.plot<-function(x){
    n=dim(x)[[1]]       # number of observations
    vp=dim(x)[[2]]      # number of variables
    S=cov(x)
 d=mahalanobis(x, colMeans(x), S)  # Mahalanobis distance^2
 d2=sort(d)
 tt=seq(1,n)
 t= (tt-0.5)/n
 q=qchisq(t, vp)
 plot(q,d2, pch="*", main="Chisquare plot for multivariate normality",
     xlab="chi2 quantile", ylab="ordered Mahalanobis d^2")
 abline(0,1)
 return(list(xbar, S))
}
chi2.plot(b)    # 그림 4.5
```

4.5.3 왜도와 첨도를 이용한 정규성 평가

▶ 정규분포 형태의 특성상 왜도(skewness)와 첨도(kurtosis)를 이용하여 정규성에 대한 평가를 할 수 있다.

n개의 확률표본 $X_1, X_2, \cdots, X_n$에 대해, 일변량 변수에 대한 왜도와 첨도

■ 왜도 : $\beta_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ (4.33)

■ 첨도 : $\beta_2 = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ (4.34)

정규분포를 따른다면 $\beta_1 = 0$, $\beta_2 = 3$이 된다.

▶ 다변량인 경우로 확장 (Mardia, 1970): 다변량 왜도와 첨도를 각각

■ 다변량 왜도 : $\beta_{1,p} = E[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^3$ (4.35)

■ 다변량 첨도 : $\beta_{2,p} = E[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^2$ (4.36)

다변량 정규성 검정에 이용.

▶ p-변량 확률벡터가 p-변량 정규분포를 따른다면

$$\beta_{1,p} = 0, \quad \beta_{2,p} = p(p+2)\ . \tag{4.37}$$

$\beta_{1,p}$와 $\beta_{2,p}$를 추정하기 위해 g_{ij}를 다음과 같이

$$g_{ij} = (\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \hat{\Sigma}^{-1} (\boldsymbol{X}_j - \overline{\boldsymbol{X}}) \tag{4.38}$$

정의한다. 여기서 $\hat{\Sigma} = \dfrac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})'$: Σ 에 대한 최대우도추정량. ▶

$\beta_{1,p}$와 $\beta_{2,p}$의 추정량으로

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3\ , \tag{4.39}$$

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2 \tag{4.40}$$

▶ 부록[표 A.6]: $p = 2, 3, 4$인 경우, $b_{1,p}$와 $b_{2,p}$ 정규성검정에 대한 기각값.
$p \geqq 5$이거나 $n \geqq 50$일 때는

■ $b_{1,p}$의 경우

$$z_1 = \frac{(p+1)(n+1)(n+3)}{6[(n+1)(p+1)-6]} b_{1,p} \tag{4.41}$$

가 근사적으로 자유도가 $\frac{1}{6}p(p+1)(p+2)$인 카이제곱분포를 따른다.

유의수준 α에서 검정법은 $z_1 \geq \chi^2_{p(p+1)(p+2)/6}(\alpha)$이면 H_0를 기각한다.

■ $b_{2,p}$의 경우

$$z_2 = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \tag{4.42}$$

근사적으로 $N(0,1)$. 유의수준 α에서 검정법은 $|z_2| \geq z_{\alpha/2}$이면 H_0를 기각한다.

≪예제 4.5≫ [표 4.1] 자료의 두 개 변수에 대해 이변량 정규분포를 따른다고 할 수 있는지를 이변량 왜도와 첨도를 계산하여 검정하고자 한다.

■ 변수가 두 개이므로 $p=2$이며 (나이, 수입)에 대한 왜도와 첨도는 각각

$$b_{1,p}=1.4117,\ \ b_{2,p}=6.2528$$

■ 유의수준 $\alpha=0.05$에서 $b_{1,p}$에 대한 기각값 3.694를 이용하면 $b_{1,p}=1.4117\ \not\geq 3.694$이므로 H_0를 기각하지 못한다.

■ 유의수준 $\alpha=0.05$에서 $b_{2,p}$에 대한 기각값 9.203을 이용하면 $b_{2,p}=6.2528\ \not\geq 9.203$이므로 H_0를 기각하지 못하므로 정규성을 벗어난다고 할 수 없다.

4.6 변수변환을 통한 정규성 만족

■ Box-Cox 변환

Box와 Cox(1964)는 $X > 0$인 경우 다음과 같은 변환식

$$X^{(\lambda)} = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln X & \lambda = 0 \end{cases}$$

을 이용해 정규성을 만족.

■ 다변량의 경우에는 확률벡터를 이루는 각 변수에 대해

$\lambda_j,\ \ j=1,2,\ \ldots\ ,p$를 구하여 $\boldsymbol{X}_i^{(\lambda)}$가 근사적으로 정규성을 갖도록 해야한다.

■ 적절한 λ_j는

$$l_j(\lambda)=-\frac{n}{2}\ln\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^{(\lambda_j)}-\overline{X_j^{(\lambda_j)}})^2\right]+(\lambda_j-1)\sum_{i=1}^{n}\ln X_i,\qquad j=1,2,\ \ldots\ ,p$$

를 최대화하는 상수. $_j$번째 관측벡터에 대한 변환식은

$$\boldsymbol{X}_i^{(\lambda)}=\begin{pmatrix}\dfrac{X_{i1}^{\lambda_1}-1}{\lambda_1}\\ \dfrac{X_{i2}^{\lambda_2}-1}{\lambda_2}\\ \vdots \\ \dfrac{X_{ip}^{\lambda_p}-1}{\lambda_p}\end{pmatrix},\qquad i=1,2,\ \ldots\ ,n$$

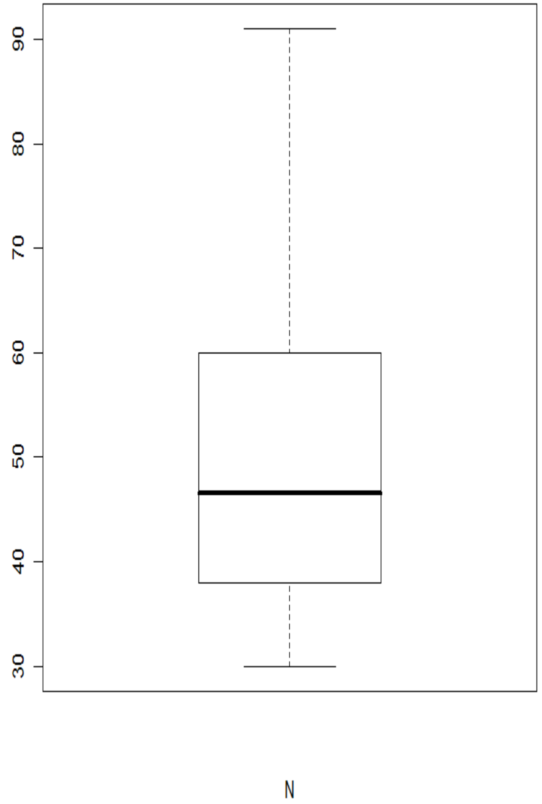
4.7 R 이용한 정규성 검토

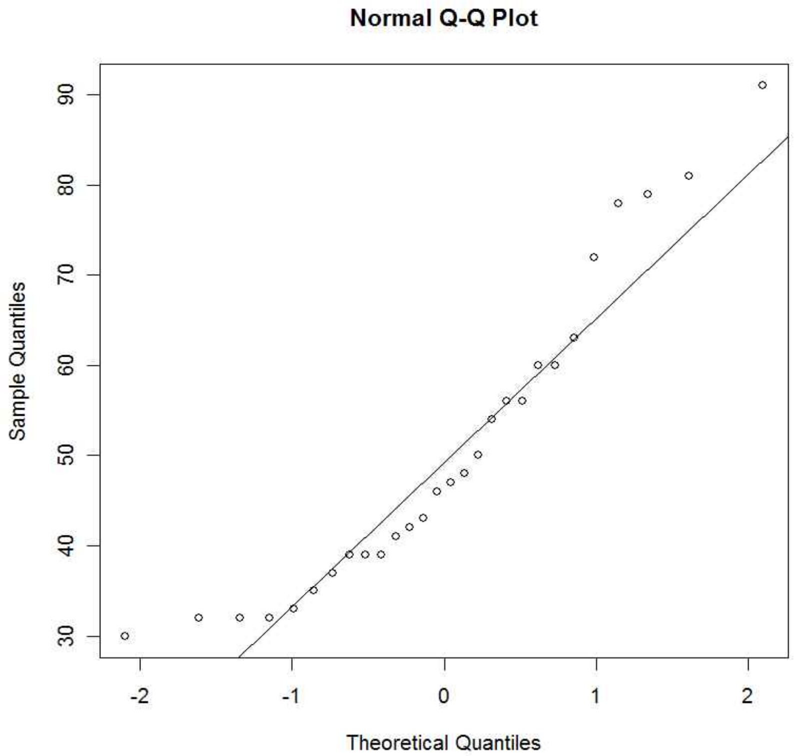
≪예제 4.6≫ 28 그루 나무를 대상으로 북쪽 동쪽 남쪽 서쪽 방향의 코르크 보오링(depths of cork boring)의 길이를 측정하여 자료를 이용하여, 각 변수가 일변량 정규분포를 따르는지를 정규확률그림(normal probability plot)을 통하여 검토하고 4개 변수가 4변량 정규분포를 따르는 지를 카이제곱그림(chi-square plot)을 그려 알아보고자 한다.

● **summary()** 함수를 이용하여 각 변수의 기초통계량을 얻을 수 있으며 **stem()** 함수로 줄기-잎 그림(stem and leaf plot)을, **boxplot()** 함수로 상자 그림을 그린다.

앞 절에서 정의한 **chi2.plot()** 함수로 카이제곱그림을 그린다.

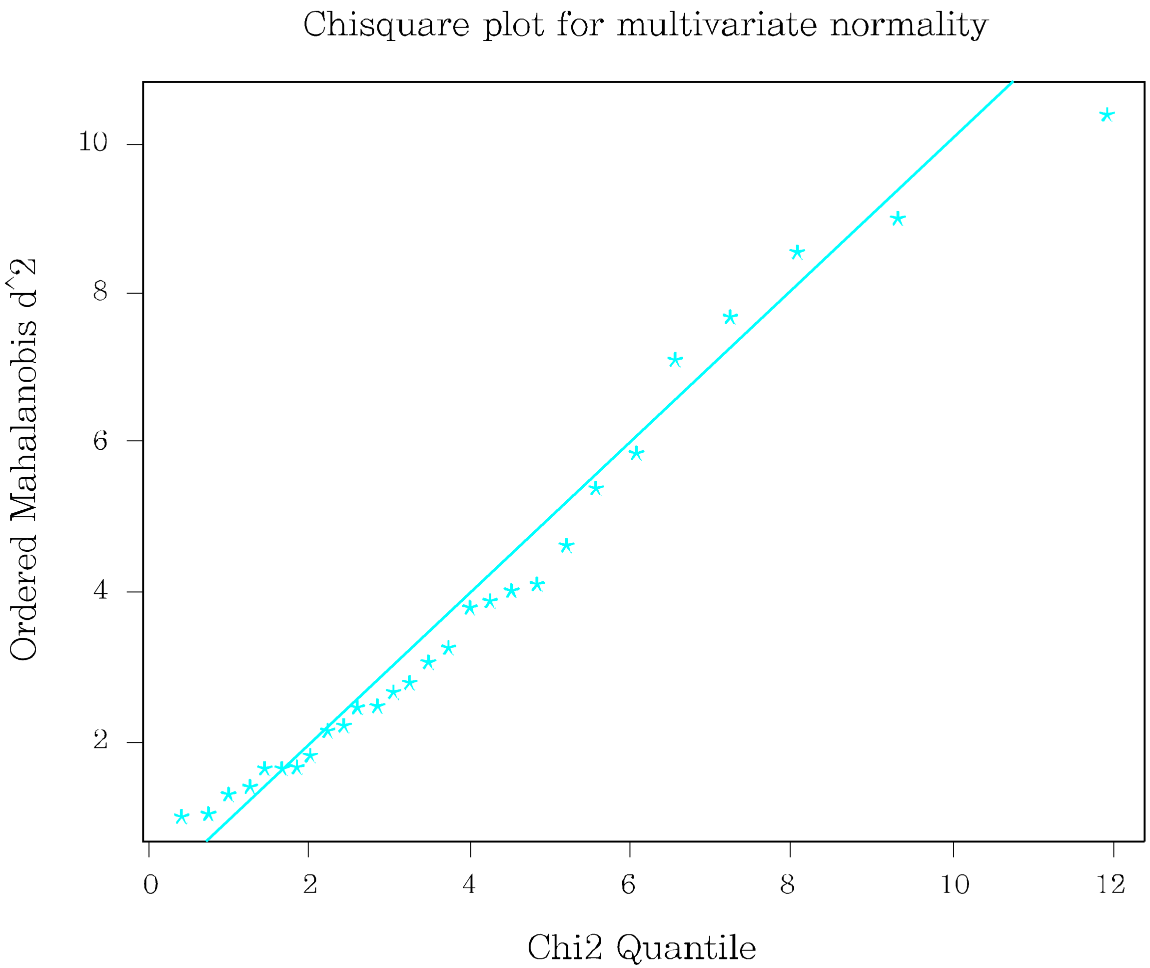
```
> summary(cork)
      N               E               S                W
 Min.   :30.00   Min.   :29.00   Min.   : 27.00   Min.   :25.00
 1st Qu.:38.50   1st Qu.:35.00   1st Qu.: 34.00   1st Qu.:35.00
 Median :46.50   Median :41.50   Median : 43.00   Median :41.50
 Mean   :50.54   Mean   :46.18   Mean   : 49.68   Mean   :45.21
 3rd Qu.:60.00   3rd Qu.:54.25   3rd Qu.: 66.25   3rd Qu.:58.00
 Max.   :91.00   Max.   :80.00   Max.   :100.00   Max.   :77.00

> stem(tree$N)
  The decimal point is 1 digit(s) to the right of the ¦
  3 ¦ 0222357999
  4 ¦ 123678
  5 ¦ 0466
  6 ¦ 003
  7 ¦ 289
  8 ¦ 1
  9 ¦ 1
> boxplot(tree$N, xlab="N")
> qqnorm(tree$N)
> qqline(tree$N)
> chi2.plot(cork)
```





[그림 4.6] Cork 자료의 N에 대한 상자그림

[그림 4.7] Cork 자료의 N에 대한 Q-Q그림



[그림 4.8] Cork 자료에 대한 카이제곱그림