응용통계연구 제 17권 1호, 2004년, pp. 35-47

다변량 정규성검정을 위한 근사 SHAPIRO-WILK 통계량의 일반화 *

김남현[1)]

요 약

본 논문에서는 Kim & Bickel(2003)에서 제안한 이변량 정규분포를 위한 검정통계량을 Fattorini(1986)의 방법을 이용하여 이변량 이상인 경우에도 실제적으로 사용가능하도록 일반화하였다. Fattorini(1986)의 통계량은 Shapiro & Wilk(1965)의 일변량 정규분포를 위한 검정통계량을 다변량으로 확장한 것이다. 그리고 제안된 통계량은 Fattorini(1986) 통계량의 근사통계량으로 생각할 수 있으며 표본의 크기가 클 때도 사용가능하다. 또한 모의실험을 통하여 여러가지 대립가설에서 기존의 통계량과의 검정력을 비교하였다.

주요용어: 다변량 정규분포, Shapiro-Wilk 통계량, 불변성.

1. 서론

$\boldsymbol{X}_1, ..., \boldsymbol{X}_n$을 $d-$차원 다변량 확률변수 $\boldsymbol{X}$의 분포에서 관측한 확률표본이라고 하자. 여기서 d는 $d \geq 1$인 고정된 정수이다. 또한 평균이 $\boldsymbol{\mu}$이고 공분산 행렬이 $\boldsymbol{\Sigma}$인 $d-$차원 다변량 정규분포를 $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$라고 하자. 대부분의 다변량 해석기법은 다변량 정규분포의 가정, 즉

H_d: $\boldsymbol{X}$의 분포가 어떤 $\boldsymbol{\mu}$와 정칙행렬 $\boldsymbol{\Sigma}$에 대해서 $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따른다.

에서 여러가지 추론방법을 제안하고 있으므로 다변량 정규분포에 대한 적합도 검정은 그 중요성을 무시할 수 없다. 따라서 다변량 정규분포를 검정하기 위한 많은 통계량들이 제안되어 온 것은 매우 당연한 일이다.

일변량 왜도와 첨도를 다변량으로 확장하는 방법이 Mardia(1970, 1974)와 Malkovich & Afifi(1973)에 의해서 제안되었다. Malkovich & Afifi(1973)은 또한 Shapiro & Wilk(1965)가 제안한 일변량 정규분포의 검정통계량을 Roy(1953)의 union-intersection 원리를 이용하여 다변량으로 확장하였다. 이는 $\boldsymbol{X}$가 다변량 정규분포를 따르면 모든 $\boldsymbol{c} \neq \boldsymbol{0}$에 대해서 $\boldsymbol{c}'\boldsymbol{X}$가 일변량 정규분포를 따른다는 사실을 이용하는 것이다. 또한 Fattorini(1986)는 Malkovich & Afifi(1973)의 통계량을 수정, 보완하였다. 그 이외의 다변량 정규성 검정에 대한 일반적인 방법에 대해서는 Mardia(1980), Thode(2002, Chapter 9) 그리고 D'Agostino & Stephens(1986, section 9.7) 등을 참고로 한다.

* 본 연구는 한국과학재단 목적기초연구(R04-2002-000-20014-0)지원으로 수행되었음.

1) (121-791) 서울시 마포구 상수동 72-1, 홍익대학교 기초과학과, 부교수
E-mail: nhkim@hongik.ac.kr

Kim & Bickel(2003)에서는 Shapiro & Wilk(1965)의 검정통계량과 밀접한 관련이 있고, 같은 극한분포를 갖는 de Wet & Vener(1972)의 일변량 정규분포의 검정통계량을 Malkovich & Afifi(1973)에서와 마찬가지로 Roy의 union-intersection 원리를 이용하여 이변량으로 일반화하였다. 또한 제안된 통계량의 귀무가설에서의 극한분포를 가우스 과정(Gaussian process)의 적분의 형태로 표현하고 모의실험을 통하여 다른 통계량과의 검정력을 비교하였다.

Kim & Bickel(2003)의 통계량은 이변량에서 $d > 2$인 $d-$변량으로의 일반화가 가능하나 이 경우 통계량의 계산이 실제적으로 용이하지 않다는 단점을 드러낸다. 본 논문에서는 Fattorini(1986)가 제안한 방법을 적용하여 Kim & Bickel(2003)의 통계량을 수정, 보완하고자 한다. 그 결과 제안된 통계량은 $d \geq 2$인 임의의 $d-$변량에서 사용가능하게 된다.

2절에서는 Malkovich & Afifi의 방법과 Fattorini의 방법을 소개하고 3절에서는 Kim & Bickel(2003)의 P_n-통계량에 Fattorini의 방법을 적용하여 얻어진 근사통계량을 제안한다. 3절에서는 제안된 통계량의 검정력을 모의실험을 통해 기존의 통계량과 비교한다.

2. Malkovich & Afifi(MA)와 Fattorini(FA)의 검정

일변량 정규분포의 검정을 위한 Shapiro-Wilk의 통계량(Shapiro & Wilk(1965)) W는

$$W(Z_1, ..., Z_n) = \frac{[\sum a_j (Z_{(j)} - \bar{Z})]^2}{\sum (Z_j - \bar{Z})^2}, \quad n \leq 50, \tag{2.1}$$

이다. 여기서 $Z_{(1)}, \ldots, Z_{(n)}$은 일변량 확률표본 $Z_1, \ldots, Z_n$의 순서통계량, $\bar{Z}$는 표본평균이고 a_j는 Shapiro & Wilk(1965)에 주어진 상수이다. Malkovich & Afifi(1973)이 제안한 방법은 적절한 상수 K_w에 대해서

$$\min_{\boldsymbol{c}} W(\boldsymbol{c}) \equiv \min_{\boldsymbol{c}} W(\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n) \geq K_w \tag{2.2}$$

이면 다변량 정규분포의 가정을 채택하는 것이다. 식(2.2)의 최소화를 위하여 MA는 $\boldsymbol{c}$가 조건

$$\boldsymbol{c}'(\boldsymbol{X}_l - \bar{\boldsymbol{X}}) = \frac{n-1}{n}, \quad \boldsymbol{c}'(\boldsymbol{X}_j - \bar{\boldsymbol{X}}) = -\frac{1}{n}, \quad j = 1, ..., n, \; j \neq l, \tag{2.3}$$

을 만족할 때 $W(\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n)$이 최소가 된다(Shapiro & Wilk(1965, Lemma 3))는 사실을 이용하여 근사해를 구하는 방법을 제안하였다. 여기서 $\bar{\boldsymbol{X}}$는 표본평균벡터이다. 식(2.3)을 만족하는 $\boldsymbol{c}$는 $n > d+1$일 때 존재하지 않으므로 MA는 최소제곱법을 이용하여 근사해를 구하는 방법을 제안하였다. 즉,

$$\left[\boldsymbol{c}'(\boldsymbol{X}_l - \bar{\boldsymbol{X}}) - \frac{n-1}{n}\right]^2 + \sum_{j \neq l}^{n} \left[\boldsymbol{c}'(\boldsymbol{X}_j - \bar{\boldsymbol{X}}) + \frac{1}{n}\right]^2$$

을 최소화하는 벡터 $\boldsymbol{c}$를 제안하였고 이는

$$\boldsymbol{c}^{(l)} = \boldsymbol{A}^{-1}(\boldsymbol{X}_l - \bar{\boldsymbol{X}}) \tag{2.4}$$

임을 쉽게 알 수 있다. 여기서

$$\boldsymbol{A} = \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$$

이다. l은 $\{1, \ldots, n\}$에서의 임의의 정수이므로 n개의 최소제곱해 $\boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(n)}$이 존재한다. MA는 $W(\boldsymbol{c})$의 분모가 최대가 되는 $\boldsymbol{c}^{(m)} \in \{\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(n)}\}$, 즉

$$(\boldsymbol{X}_m - \bar{\boldsymbol{X}})' \boldsymbol{A}^{-1} (\boldsymbol{X}_m - \bar{\boldsymbol{X}}) = \max_{1 \le l \le n} (\boldsymbol{X}_l - \bar{\boldsymbol{X}})' \boldsymbol{A}^{-1} (\boldsymbol{X}_l - \bar{\boldsymbol{X}})$$

을 만족하는 $\boldsymbol{c}^{(m)}$을 택하였다. 따라서 MA 통계량은

$$\begin{aligned} MA(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) &= W(\boldsymbol{c}^{(m)}) = W(\boldsymbol{c}^{(m)'} \boldsymbol{X}_1, ..., \boldsymbol{c}^{(m)'} \boldsymbol{X}_n) \\ &= \frac{[\sum_{j=1}^{n} a_j (U_{(j)} - \bar{U})]^2}{(\boldsymbol{X}_m - \bar{\boldsymbol{X}})' \boldsymbol{A}^{-1} (\boldsymbol{X}_m - \bar{\boldsymbol{X}})} \end{aligned}$$

이다. 여기서 $U_{(1)} \le \cdots \le U_{(n)}$은 $U_j = (\boldsymbol{X}_m - \bar{\boldsymbol{X}})' \boldsymbol{A}^{-1} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})$, $j = 1, ..., n$의 순서통계량이다.

한편, $MA(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$은 n개의 가능한 해인 $\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(n)}$에 대해서조차 $W(\boldsymbol{c})$을 최소화하지 못한다는데 착안하여 Fattorini(1986)는

$$\begin{aligned} FA(\boldsymbol{X}_1, ..., \boldsymbol{X}_n) &= \min_{1 \le l \le n} W(\boldsymbol{c}^{(l)}) \\ &= \min_{1 \le l \le n} \frac{[\sum_{j=1}^{n} a_j (U_{(j)} - \bar{U})]^2}{(\boldsymbol{X}_l - \bar{\boldsymbol{X}})' \boldsymbol{A}^{-1} (\boldsymbol{X}_l - \bar{\boldsymbol{X}})} \end{aligned} \tag{2.5}$$

을 제안하였다. 여기서 $U_{(j)}$는

$$U_j = (\boldsymbol{X}_l - \bar{\boldsymbol{X}})' \mathbf{A}^{-1} (\boldsymbol{X}_j - \bar{\boldsymbol{X}}), \quad j = 1, ..., n, \tag{2.6}$$

의 순서통계량이다. 당연히 $FA \le MA$가 성립하고 두 통계량 모두 벡터합과 정칙행렬곱에 대해서 불변이다. 또한 상수 a_j, $j = 1, \ldots, n$은 $n \le 50$일 때 Shapiro & Wilk(1965)에 주어져 있으므로, MA와 FA는 $n \le 50$일 때 사용가능하다.

3. 제안된 검정통계량

일변량 정규분포의 검정을 위한 de Wet & Venter(1972)의 통계량은

$$L_n(Z_1, \ldots, Z_n) = \sum_{i=1}^{n} \left(\frac{Z_{(i)} - \bar{Z}}{s} - H_i \right)^2 \tag{3.1}$$

이다. 여기서 s^2은 표본분산 $s^2 = n^{-1} \sum (Z_i - \bar{Z})^2$이고 $H_i = \Phi^{-1}(\frac{i}{n+1})$, Φ^{-1}는 표준정규분포 $N_1(0, 1)$의 분포함수 Φ의 역함수이다.

de Wet & Venter(1972)의 L_n-통계량은 일변량 정규성 검정을 위한 Shapiro & Wilk(1965)의 W-통계량, Shapiro & Francia(1972)의 W'-통계량과 밀접한 관련이 있다. 사실상 L_n-통계량은 W-통계량의 간단한 형태로 생각될 수 있고 (D'Agostino & Stephens(1986, 5.10절), de Wet & Venter(1972)), 세 통계량은 모두 같은 근사분포를 갖는다는 것이 증명되었다(Leslie et al.(1986)). 이들 통계량의 극한분포에 대해서는 de Wet & Venter(1972, 1973), Csörgő(1983, 7장), del Barrio et al.(1999) 등을 참고로 한다.

$$r_n(Z_1,\ldots,Z_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_{(i)}H_i}{st}, \quad t^2 = \frac{1}{n}\sum_{i=1}^{n}H_i^2 \tag{3.2}$$

이라고 하면, 일반적으로 $r_n > 0$이고

$$L_n = 2nt(1-r_n) + n(1-t)^2 \tag{3.3}$$

임을 쉽게 보일 수 있다.

식 (3.2)의 r_n의 제곱은

$$r_n^2 = \frac{[\sum d_j(Z_{(j)} - \bar{Z})]^2}{\sum(Z_j - \bar{Z})^2}$$

으로 쓸 수 있다. 여기서

$$\boldsymbol{d} = (d_1,\ldots,d_n)' = \boldsymbol{H}/(\boldsymbol{H}'\boldsymbol{H})^{1/2}, \quad \boldsymbol{H} = (H_1,\ldots,H_n)'$$

이다. 따라서 $\sum d_i = 0$이 성립한다. 즉 r_n^2은 식 (2.1)의 W-통계량과 계수 a_j를 제외하고 같은 형태로 표현된다. 따라서 Shapiro & Wilk(1965)의 Lemma 3과 유사한 다음의 보조정리를 얻을 수 있다.

보조정리 3.1 r_n^2은 최소값 $nd_1^2/(n-1)$을 갖는다.

증명: Shapiro & Wilk(1965)의 Lemma 3의 증명과 같은 방법을 적용한다. r_n^2은 위치, 척도 불변인 통계량이므로 $\sum Z_i = 0$, $\sum d_i Z_{(i)} = 1$이라는 제한조건에서 $\sum Z_i^2$의 최대값을 고려하면 충분하다. 이것은 볼록영역(convex region)이고 $\sum Z_i^2$도 볼록함수(convex function)이므로 C. L. Mallows에 의해 최대값은 다음 $n-1$개의 점에서 나타난다.

$$\begin{aligned}
\boldsymbol{p}_1 &= \left(\frac{n-1}{nd_1}, \frac{-1}{nd_1}, \ldots, \frac{-1}{nd_1}\right) \\
\boldsymbol{p}_2 &= \left(\frac{n-2}{n(d_1+d_2)}, \frac{n-2}{n(d_1+d_2)}, \frac{-2}{n(d_1+d_2)}, \ldots, \frac{-2}{n(d_1+d_2)}\right) \\
&\vdots \\
\boldsymbol{p}_{n-1} &= \left(\frac{1}{n(d_1+\cdots+d_{n-1})}, \frac{1}{n(d_1+\cdots+d_{n-1})}, \cdots, \frac{-(n-1)}{n(d_1+\cdots+d_{n-1})}\right)
\end{aligned}$$

주어진 계수 $\boldsymbol{d}$에 대해서 $\sum Z_i^2$의 최대값이 $\boldsymbol{p_1}$(또는 $\boldsymbol{p}_{n-1}$)에서 나타나고 해당하는 r_n^2의 최소값이 $nd_1^2/(n-1)$이라는 것을 수치적으로 보일 수 있다.

다시 말하면, 각 $\boldsymbol{p}_k$에서 $\sum Z_i^2 = k(n-k)/n(d_1+\cdots+d_k)^2 \equiv \alpha(k)$이고, $d_i = -d_{n-i+1}$, $\sum d_i = 0$이므로 $\alpha(k) = \alpha(n-k)$이 성립한다. 따라서 $k = 1, \ldots, [n/2]$을 고려하면 충분하다. 그 결과 $\alpha(k)$가 $k = 1$일 때 최대가 된다는 것을 보이기 위해서

$$\frac{n-1}{n}\frac{1}{d_1^2} \geq \frac{k(n-k)}{n}\frac{1}{(d_1+\cdots+d_k)^2}, \quad k = 1, \ldots, [n/2]$$
$$\frac{d_1^2}{(d_1+\cdots+d_k)^2} \leq \frac{n-1}{k(n-k)}, \quad k = 1, \ldots, [n/2]$$

임을 보이면 충분하고 이는 주어진 $\boldsymbol{d} = (d_1, \ldots, d_n)'$에 대해서 수치적으로 쉽게 보일 수 있다. □

Kim & Bickel(2003)에서는 $d = 2$, $\boldsymbol{X} = (X_1, X_2)'$일때 복합귀무가설 H_2를 검정하기 위하여

$$P_n = \max_{c_1,c_2} \sum_{i=1}^{n} \left\{ \frac{(c_1X_1 + c_2X_2)_{(i)} - (c_1\bar{X}_1 + c_1\bar{X}_1)}{sd(c_1X_1 + c_2X_2)} - H_i \right\}^2 \tag{3.4}$$

을 제안하였다. 여기서 $\bar{X}_k = \frac{1}{n}\sum_{i=1}^n X_{ki}$, $sd^2(c_1X_1 + c_2X_2) = c_1^2\hat{\sigma}_1^2 + c_2^2\hat{\sigma}_2^2 + 2c_1c_2\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2$, $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^n (X_{ki} - \bar{X}_k)^2$, $k = 1, 2$, $\hat{\rho} = \frac{1}{n}\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)/(\hat{\sigma}_1\hat{\sigma}_2)$이고 $(\cdot)_{(i)}$는 괄호 안의 확률변수의 i번째 순서통계량이다. P_n-통계량은 식(3.1)의 L_n-통계량을 Roy의 union-intersection 원리를 이용하여 이변량으로 일반화한 것이다. 그리고 P_n-통계량은 벡터합과 정칙행렬곱에 대해서 불변(invariance)이다. 따라서 $(c_1, c_2) = (\cos\theta, \sin\theta)$, $0 \leq \theta \leq \pi$로 가정해도 무방하다. 즉, 이변량 분포일 때는 P_n-통계량을 계산할 때 단일변수 θ에서 최대값을 고려하면 충분하다.

식(3.4)의 P_n-통계량은 벡터를 이용하여 표현하면

$$P_n = \max_{\boldsymbol{c}} \sum_{i=1}^{n} \left[\frac{(\boldsymbol{c}'(\boldsymbol{X} - \bar{\boldsymbol{X}}))_{(i)}}{(\frac{1}{n}\boldsymbol{c}'A\boldsymbol{c})^{1/2}} - H_i \right]^2$$

이다. 즉, 통계량 P_n은 이변량뿐만 아니라 $d > 2$인 다변량에서도 같은 방법으로 정의될 수 있다. 그러나 $d > 2$인 다변량에서는 $d = 2$인 경우와 달리, P_n의 계산이 실제적으로 용이하지 않다. 이 절에서는 이러한 P_n의 단점을 해결하기 위해서 P_n의 근사통계량을 Fattorini(1986)의 방법을 이용하여 제안하고자 한다.

식(3.3)에 의해서

$$P_n = 2nt(1 - \min_{\boldsymbol{c}} r_n(\boldsymbol{c})) + n(1-t)^2 \tag{3.5}$$

이다. 여기서 $r_n^2(\boldsymbol{c}) = r_n^2(\boldsymbol{c}'\boldsymbol{X}_1, \ldots, \boldsymbol{c}'\boldsymbol{X}_n)$을 의미한다. 일반적으로 $r_n(\boldsymbol{c}) > 0$이므로

$$R_n^2 \equiv \min_{\boldsymbol{c}} r_n^2(\boldsymbol{c})$$

라고 하면 P_n이 클때 귀무가설 H_d를 기각하는 검정은 R_n^2이 작을 때 H_d를 기각하는 검정과 동일하다.

보조정리 3.1을 근거로, R_n^2 역시 $\boldsymbol{c}$가 식(2.3)의 조건을 만족할 때 최소가 되고 최소제곱법을 이용한 근사해는 식(2.4)의 $\boldsymbol{c}^{(l)}$로 주어진다. 따라서 R_n^2 또는 P_n에 Fattorini의 방법을 적용하면 다음과 같은 근사통계량

$$R_n^{2*} = \min_{1\le l\le n} \frac{[\sum_{j=1}^n d_j(U_{(j)}-\bar{U})]^2}{(\boldsymbol{X}_l-\bar{\boldsymbol{X}})'\boldsymbol{A}^{-1}(\boldsymbol{X}_l-\bar{\boldsymbol{X}})} \tag{3.6}$$

$$\begin{aligned} P_n^* &= \max_{1\le l\le n} L_n(c^{(l)}) \\ &= \max_{1\le l\le n} \sum_{j=1}^n \left(\frac{U_{(j)}-\bar{U}}{(\frac{1}{n}(\boldsymbol{X}_l-\bar{\boldsymbol{X}})'\boldsymbol{A}^{-1}(\boldsymbol{X}_l-\bar{\boldsymbol{X}}))^{1/2}} - H_j \right)^2 \end{aligned} \tag{3.7}$$

을 얻을 수 있다. 여기서 $U_{(j)}$는 식(2.6)의 U_j의 순서통계량이다.

식(3.4)의 P_n-통계량과 식(3.7)의 P_n^*-통계량의 근사정도를 보기 위하여 모의실험을 행하였다. 이변량 정규분포 $N_2(\boldsymbol{0},\boldsymbol{I})$에서 표본크기 $n = 10(10)50, 100$인 표본 $N = 1000$개를 추출하여 상대오차

$$D = \frac{P_n - P_n^*}{P_n}$$

을 구하여 각 표본크기에서의 평균을 표 3.1과 그림 3.1에 제시하였다. 이로부터 표본크기가 커짐에 따라 상대오차가 현저하게 감소함을 볼 수 있고 따라서 P_n^*는 P_n의 합리적인 근사통계량이라고 볼 수 있다.

표 3.1: 표본크기 $n = 10(10)50, 100$인 $N = 1000$개의 표본에서 계산된 상대오차의 평균

n	10	20	30	40	50	100
상대오차평균	0.06805	0.03598	0.02247	0.01667	0.01209	0.004225

식(3.4) 또는 식(3.7)의 $H_i = \Phi^{-1}(\frac{i}{n+1})$는 잘 알려진 바와 같이 $m_i \equiv E(Z_{(i)})$의 근사값으로 생각할 수 있다. 여기서 $Z_{(1)},\ldots,Z_{(n)}$은 $N_1(0,1)$에서의 확률표본 $Z_1,\ldots,Z_n$의 순서통계량이다. 따라서 n이 크지 않을 때는 H_i 대신 m_i나 이의 근사값인 $\Phi^{-1}\left(\frac{i-0.375}{n+0.125}\right)$를 이용하는 것이 좀 더 합리적이라고 생각된다. 이러한 근사는 Blom(1958)이 처음으로 제안하였고, 표준정규분포에서의 순서통계량 m_i는 수치계산을 통하여 얻어지며 Harter(1961)에 $n \le 400$일 때 주어져 있다. 따라서 식(3.7)의 P_n^*에 H_i 대신 m_i를 대입한 통계량, 또는 이와 동일하게(식(3.5) 참조) 식(3.6) 대신

$$R_n^{2**} = \min_{1\le l\le n} \frac{[\sum_{j=1}^n b_j(U_{(j)}-\bar{U})]^2}{(\boldsymbol{X}_l-\bar{\boldsymbol{X}})'\boldsymbol{A}^{-1}(\boldsymbol{X}_l-\bar{\boldsymbol{X}})}, \tag{3.8}$$

$\boldsymbol{b} = (b_1,\ldots,b_n)' = \boldsymbol{m}/(\boldsymbol{m}'\boldsymbol{m})^{1/2}$, $\boldsymbol{m} = (m_1,\ldots,m_n)' = (E(Z_{(1)}),\ldots,E(Z_{(n)}))'$을 고려할 수 있다.

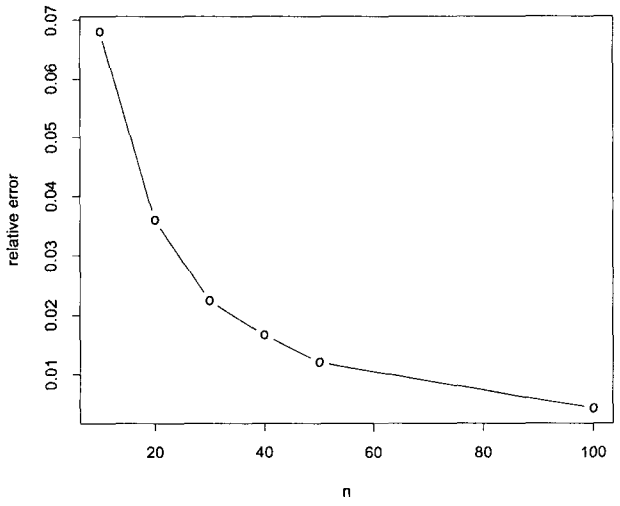

그림 3.1: 표본크기 n과 상대오차의 plot

식(3.8)의 R_n^{2**}은 Shapiro & Francia(1972)가 제안한 W'-통계량을 Fattorini의 방법을 이용해 다변량으로 확장한 통계량이라고 할 수 있다. 따라서 R_n^{2**}은 $n > 50$일 때도 사용가능한, 식(2.5)의 FA-통계량의 근사통계량이라고 생각할 수 있다.

4. 모의실험결과

3절에서 언급한 이유로 n이 크지 않을 때는 식(3.7)의 P_n^*보다 식(3.8)의 R_n^{2**}를 이용하는 것이 좀 더 합리적이라고 생각되므로 이 절에서는 R_n^{2**}-통계량의 근사백분위수와 몇개의 대립가설에서의 검정력을 알아보기 위한 모의실험을 행하였다.

표본크기 $n = 10(10)50, 100$에서 $N = 5000$개의 표본을 평균이 $\mathbf{0}$이고 공분산 행렬 $\boldsymbol{\Sigma} = \boldsymbol{I}$인 정규분포로부터 추출하여 R_n^{2**}의 값을 각 표본으로부터 계산하였다. 표 4.1와 표 4.2에는 각각 $d = 2$일때와 $d = 5$일때의 유의수준 α에서의 기각값이 제시되어있다. R_n^{2**}은 벡터합과 정칙행렬곱에 의해서 불변인 통계량이므로 주어진 기각값은 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 무관하다.

다음으로 R_n^{2**}의 검정력을 표본크기 $n = 20$, $n = 50$, 유의수준 $\alpha = 0.05$에서 살펴보았다. Henze & Zirkler(1990)는 벡터합과 정칙행렬곱에 대해서 불변인 몇 가지 다변량 정규분포를 위한 검정통계량의 검정력을 비교하였고 MA와 FA도 비교 대상에 포함하였다. 그들은 (i) 주변분포가 서로 독립인 분포 (ii) 혼합정규분포(mixtures of normal distributions) 등을 고려하였다. 표 4.3, 4.4에서 $N(0,1)$, $C(0,1)$, $Logis(0,1)$, exp(1)은 각각 표준정규분포, 코쉬분포, 로지스틱분포, 지수분포를 나타낸다. χ_k^2과 t_k는 자유도가 k인 카이제곱분포와 t분포를 나타낸다. $\Gamma(a,b)$는 확률밀도함수가

$$b^{-a}\Gamma(a)^{-1}x^{a-1}\exp(-x/b),\ x > 0$$

인 감마분포이고 $B(a,b)$는 확률밀도함수

$$B(a,b)^{-1}x^{a-1}(1-x)^{b-1},\ 0<x<1,$$

인 베타분포이고 $LN(a,b)$는 확률밀도함수

$$(\sqrt{2\pi}bx)^{-1}\exp(-(\log x-a)^2/2b^2),\ x>0,$$

인 대수정규분포를 나타낸다. 또한 $F_1 * F_2$는 서로독립인 주변분포 F_1과 F_2를 갖는 분포이며 F_1^2은 각각의 주변분포가 서로독립인 F_1분포임을 의미한다. $NMIX_2(\kappa,\delta,\rho_1,\rho_2)$는

$$\kappa N_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho_1\\\rho_1&1\end{pmatrix}\right)+(1-\kappa)N_2\left(\begin{pmatrix}\delta\\\delta\end{pmatrix},\begin{pmatrix}1&\rho_2\\\rho_2&1\end{pmatrix}\right)$$

인 이변량 혼합정규분포를 말한다.

표 4.1: R_n^{2**}-통계량의 근사백분위수 ($n=10(10)50,100$, $d=2$)

n	α =0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
10	0.6809	0.7139	0.7559	0.7946	0.8782	0.9238	0.9318	0.9395	0.9437
20	0.8031	0.8223	0.8504	0.8727	0.9246	0.9507	0.9550	0.9604	0.9626
30	0.8531	0.8748	0.8938	0.9079	0.9439	0.9626	0.9660	0.9699	0.9718
40	0.8882	0.8997	0.9137	0.9260	0.9548	0.9703	0.9730	0.9753	0.9771
50	0.9084	0.9180	0.9296	0.9392	0.9626	0.9747	0.9772	0.9796	0.9809
100	0.9483	0.9543	0.9620	0.9668	0.9792	0.9858	0.9871	0.9882	0.9889

표 4.2: R_n^{2**}-통계량의 근사백분위수 ($n=10(10)50,100$, $d=5$)

n	α =0.01	0.02	0.05	0.10	0.5	0.9	0.95	0.98	0.99
10	0.4403	0.4659	0.5054	0.5426	0.6667	0.7679	0.7932	0.8115	0.8234
20	0.6622	0.6887	0.7264	0.7560	0.8380	0.8888	0.8989	0.9090	0.9145
30	0.7660	0.7879	0.8126	0.8350	0.8889	0.9219	0.9287	0.9353	0.9382
40	0.8185	0.8346	0.8580	0.8728	0.9147	0.9395	0.9445	0.9487	0.9518
50	0.8529	0.8674	0.8849	0.8971	0.9299	0.9496	0.9532	0.9572	0.9591
100	0.9221	0.9281	0.9375	0.9436	0.9616	0.9715	0.9734	0.9752	0.9762

표 4.3와 표 4.4의 검정력은 $N=1000$개의 표본 중 유의한 표본의 백분위를 소수 첫째자리에서 반올림한 것이다. 표 4.3와 표 4.4에서는 R_n^{2**}의 검정력을 Henze & Zirkler(1990)에

주어진 MA, FA의 검정력과 비교하고 있다. 또한 표 4.3에서는 Kim & Bickel(2003)에 주어진 P_n의 검정력과도 비교하고 있다. 표 4.3의 검정력으로부터 P_n의 근사통계량으로 제안된 R_n^{2**}의 검정력은 P_n의 검정력과 비슷한 양상을 보임을 알 수 있다. 따라서 Kim & Bickel(2003)에서 P_n-통계량의 검정력에 대해서 지적한 바와 같이 R_n^{2**}의 검정력도 주어진 대립가설에서는 일반적으로 MA보다는 우수하고 FA보다는 약간 떨어진다고 할 수 있다. 특히 주변분포가 꼬리가 짧은 분포인 $B(1,1)^2$, $B(1,2)^2, B(2,2)^2$ 또는 $N(0,1)*B(1,1)$일 경우는 FA가 매우 우수하고, 주변분포가 꼬리가 긴 분포인 $(t_5)^2$, $Logis(0,1)^2$, $N(0,1)*t_5$인 경우에는 P_n의 경우와 마찬가지로 R_n^{2**}도 FA보다 약간 검정력이 우수함을 볼 수 있다. 이는 일변량 정규분포의 검정통계량인 W와 W'에도 나타나는 현상이다(Shapiro & Francia(1972), Looney & Gulledge(1985)). 그러나 이러한 경향은 차원이 높아지면서(표 4.4) 점점 그 정도가 약해짐을 볼 수 있다.

참고문헌

[1] Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates.* Wiley, New York.

[2] Csörgő, M. (1983). *Quantile Processes with Statistical Applications.* CBMS-NSF Regional Conference Series in Applied Mathematics.

[3] D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques.* Marcel Dekker, New York.

[4] de Wet, T. and Venter, J. H. (1972). Asymptotic distributions of certain test criteria of normality. *South African Statistical Journal,* **6**, 135-149.

[5] de Wet, T. and Venter, J. H. (1973). Asymptotic distributions for quadratic forms with applications to tests of fit. *The Annals of Statistics,* **1**, 380-387.

[6] del Barrio, E., Cuesta, J. A., Matrán, C. and Rodríguez, J. M. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics,* **27**, 1230-1239.

[7] Fattorini, L. (1986). Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality. *Statistica,* **46**, 209-217.

[8] Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika,* **48**, 151-165.

[9] Henze, N. and Zirkler, H. (1990). A class of invariant and consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods,* **19**, 3595-3617.

[10] Kim, N. and Bickel, P. J. (2003). The limit distribution of a test statistic for bivariate normality. *Statistica Sinica,* **13**, 327-349.

[11] Leslie, J. R., Stephens, M. A. and Fotopolous, S. (1986). Asymptotic distribution of the

Shapiro-Wilk W for testing for normality. *The Annals of Statistics,* **14**, 1497-1506.

[12] Looney, S. W. and Gulledge, T. R. Jr. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician,* **39**, 75-79.

[13] Malkovich, J. F. and Afifi, A. A. (1973). On tests for multivariate normality. *Journal of the American Statistical Association,* **68**, 176-179.

[14] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika,* **57**, 519-530.

[15] Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhya A,* **36**, 115-128.

[16] Mardia, K. V. (1980). Tests of univariate and multivariate normality. In *Handbook in Statistics* (Ed. P. R. Krishnaiah), 279-320. Amsterdam, North-Holland.

[17] Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics,* **24**, 220-238.

[18] Shapiro, S. S. and Francia, R. S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association,* **67**, 215-216.

[19] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* **52**, 591-611.

[20] Thode, H. C. Jr. (2002). *Testing for Normality.* Marcel Dekker, New York.

[2003년 3월 접수, 2003년 7월 채택]

표 4.3: 각 분포에서 P_n, R_n^{2**}, MA, FA 통계량의 검정력 비교
(유의수준 $\alpha = 0.05$, $n = 20,\ 50$, $d = 2$)

	$n = 20$				$n = 50$			
대립가설	P_n	R_n^{2**}	MA	FA	P_n	R_n^{2**}	MA	FA
$N(0,1)^2$	5	5	5	5	5	5	5	5
$\exp(1)^2$	81	80	76	86	100	100	100	100
$LN(0,.5)^2$	54	55	53	59	94	96	92	97
$C(0,1)^2$	98	97	96	96	100	100	–	–
$\Gamma(5,1)^2$	24	21	22	25	56	59	58	67
$(\chi_5^2)^2$	37	41	43	44	87	89	84	93
$(\chi_{15}^2)^2$	18	17	18	17	42	41	42	46
$(t_2)^2$	72	70	69	68	97	97	94	95
$(t_5)^2$	28	25	24	22	55	51	46	40
$B(1,1)^2$	1	1	2	6	6	19	4	77
$B(1,2)^2$	12	10	9	19	42	54	35	86
$B(2,2)^2$	1	1	2	3	1	1	2	15
$Logis(0,1)^2$	15	15	16	15	30	29	21	16
$N(0,1) * \exp(1)$	58	56	52	63	98	98	87	99
$N(0,1) * \chi_5^2$	25	25	26	28	65	69	61	73
$N(0,1) * t_5$	16	16	16	16	34	33	24	19
$N(0,1) * B(1,1)$	3	2	4	6	4	12	4	56
$NMIX_2(.5,2,0,0)$	4	4	4	4	4	4	4	17
$NMIX_2(.5,4,0,0)$	14	21	4	51	95	98	5	100
$NMIX_2(.5,2,.9,0)$	31	23	27	29	68	65	54	66
$NMIX_2(.5,.5,.9,0)$	23	20	21	20	48	43	33	29
$NMIX_2(.5,.5,.9,-.9)$	51	48	47	51	93	91	76	83

표 4.4: 각 분포에서 R_n^{2**}, MA, FA 통계량의 검정력 비교
(유의수준 $\alpha = 0.05$, $n = 20,\ 50$, $d = 5$)

	$n = 20$			$n = 50$		
대립가설	R_n^{2**}	MA	FA	R_n^{2**}	MA	FA
$N(0,1)^5$	5	5	5	6	5	6
$\exp(1)^5$	60	61	65	100	97	100
$LN(0,.5)^5$	45	47	49	93	90	96
$C(0,1)^5$	99	99	99	100	-	-
$\Gamma(0.5,1)^5$	89	86	90	100	-	-
$\Gamma(5,1)^5$	17	15	15	48	45	54
$(\chi_5^2)^5$	29	30	31	79	73	84
$(\chi_{15}^2)^5$	13	14	15	33	31	34
$(t_2)^5$	81	80	81	100	99	100
$(t_5)^5$	25	28	29	62	55	56
$B(1,1)^5$	1	1	1	0	0	1
$B(1,2)^5$	3	4	5	4	5	14
$B(2,2)^5$	1	2	2	1	1	1
$Logis(0,1)^5$	11	13	13	30	27	27
$N(0,1)^4 * \exp(1)$	18	21	21	61	56	72
$N(0,1)^4 * \chi_5^2$	8	10	10	30	29	34
$N(0,1)^4 * t_5$	9	10	10	22	19	19

An Approximate Shapiro-Wilk Statistic for Testing Multivariate Normality *

Namhyun Kim [1)]

ABSTRACT

In this paper, we generalizes Kim and Bickel(2003)'s statistic for bivariate normality to that of multinormality, applying Fattorini(1986)'s method. Fattorini(1986) generalized Shapiro-Wilk's statistic for univariate normality to multivariate cases. The proposed statistic could be considered as an approximate statistic to Fattorini(1986)'s. It can be used even for a big sample size. Power performance of the proposed test is assessed in a Monte Carlo study.

Keywords: Multivariate normality; Shapiro-Wilk statistic; Invariance.

* This work was supported by grant No. R04-2002-000-20014-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

1) Associate Professor, Dept. of Science, Hongik University.
E-mail: nhkim@hongik.ac.kr