응용통계연구 제 17권 3호, 2004년, pp. 507-518

정규성 검정을 위한 다변량 왜도와 첨도의 이용에 대한 고찰 *

김남현[1)]

요 약

Malkovich & Afifi (1973)는 합교원리(union-intersection principle)를 이용하여 왜도와 첨도를 다변량으로 일반화하였으나 이는 자료의 차원이 클 경우에는 사용이 용이하지 않다. 본 논문에서는 이러한 단점을 보완하는 이들의 근사통계량을 제안한다. 그리고 제안된 근사통계량, Malkovich & Afifi (1973)의 통계량, Mardia(1970)의 왜도와 첨도의 검정력을 모의실험을 통하여 비교한다.

주요용어: 적합도 검정, 다변량 정규분포, 왜도, 첨도.

1. 서론

다변량 정규성에 대한 검정은 오랫동안 통계적 추론의 주요한 연구 주제 중 하나였다. 따라서 이에 대한 많은 검정방법이 제안된 것은 매우 당연한 일이다. 다변량 정규성 검정에 대한 일반적인 방법에 대해서는 Mardia(1980), D'Agostino & Stephens(1986, section 9.7), Thode(2002, Chapter 9) 그리고 Henze(2002) 등을 참고로 한다.

우선 일변량 왜도와 첨도를 다변량으로 확장하는 방법이 Mardia(1970, 1974, 1975)와 Malkovich & Afifi (1973)에 의해 제안되었다. 그리고 Baringhaus & Henze(1991, 1992), de Wet, Venter & van Wyk(1979), Machado(1983) 등은 이들의 극한분포에 대해서 연구하였다. 또한 Malkovich & Afifi (1973), Fattorini(1986), Kim & Bickel(2003), 김남현(2004) 등은 Shapiro & Wilk(1965)가 제안한 일변량 정규분포의 검정통계량 또는 이의 근사통계량을 다변량으로 확장하였다. 그리고 Baringhaus & Henze(1988), Henze & Zirkler(1990), Henze & Wagner(1997)는 empirical characteristic function을 이용한 검정방법에 대해서 연구하였고, Zhu, Fang & Bhatti(1997), Zhu, Wong & Fang(1995)은 사영추적(projection pursuit)을 이용한 검정법에 대해서 연구하였다. 또한 Horswell & Looney(1992)와 Romeu & Ozturk(1993)은 다변량 정규성 검정을 위한 여러가지 통계량을 비교연구하였다.

앞에서 언급한 바와 같이 Mardia(1970, 1974, 1975)와 Malkovich & Afifi(1973)는 각각 다른 방법으로 왜도와 첨도를 다변량으로 확장하였다. Mardia(1970)의 통계량은 자주 인용되는 검정법인 반면 Malkovich & Afifi(1973)는 이변량일 경우를 제외하고는 실제로 이용하기 불편한 형태로 정의되어있다. 그러나 두 통계량은 밀접한 관련이 있을 것이라 예상되며 이에 대한 연구가 필요하다고 생각된다. 본 논문에서는 Malkovich & Afifi(1973)가 정

* 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

의한 다변량 왜도와 첨도의 근사계산방법을 제안하고 이 근사통계량을 Malkovich & Afifi (1973)의 통계량, Mardia(1970)의 통계량과 비교해보고자 한다. 이를 위해서 여러가지 대립가설에서의 검정력을 모의실험을 통하여 살펴보았다.

2. Malkovich & Afifi(MA)의 왜도와 첨도의 근사통계량

$\boldsymbol{X}_1, ..., \boldsymbol{X}_n$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 분산-공분산 행렬이 $\boldsymbol{\Sigma}$인 $d-$차원 다변량 정규분포를 $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$라고 하자. 다변량 정규분포의 가정은

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

를 검정하는 것이다.

왜도나 첨도등의 일차적률을 이용하는 검정법이 일변량 정규성검정에서 자주 사용되며, 이들이 좋은 검정력을 보여준다는 것이 알려져 있다(Pearson, D'Agostino & Bowman(1977)). 따라서 다변량 정규성 검정을 위하여 왜도나 첨도를 다변량으로 확장하려는 시도는 매우 당연하다. Mardia(1970, 1974, 1975)와 Malkovich & Afifi(1973)는 각각 다른 방법으로 왜도와 첨도를 다변량으로 확장하였다.

Mardia는 일변량 왜도, 첨도와 관계된 t-통계량의 로버스트성에 대한 사실을 다변량으로 확장함으로써 다변량 왜도와 첨도를 정의하였다. $\bar{\boldsymbol{X}}$를 확률표본 $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$의 표본평균벡터, $\mathbf{S}$를 표본 분산-공분산행렬이라고 하자. 즉, $\mathbf{S}$는

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

이다. Mardia의 왜도는

$$b_{1,d} = \frac{1}{n^2}\sum_i\sum_j \left\{(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})\right\}^3 \tag{2.1}$$

으로 정의되고 Mardia의 첨도는

$$b_{2,d} = \frac{1}{n}\sum_i \left\{(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})\right\}^2 \tag{2.2}$$

이다. 여기서 $'$은 전치(transpose)를 의미한다. Mardia의 왜도와 첨도는 아마도 다변량 정규성 검정을 위한 통계량 중 가장 자주 인용되는 것이라 생각된다(Thode(2002, 9.4절 p.196)). 또한 통계적인 이론이나 성질이 가장 많이 연구된 방법이라고 할 수 있다(Henze(2002, 10절)).

반면 Malkovich & Afifi(1973)는 Roy(1953)의 합교원리(union-intersection principle)를 이용하여 일변량 왜도와 첨도를 다변량으로 확장하였다. 이는 $\boldsymbol{X}$가 다변량 정규분포를 따르면 모든 $\boldsymbol{c} \neq \mathbf{0}$에 대해서 $\boldsymbol{c}'\boldsymbol{X}$가 일변량 정규분포를 따른다는 사실을 이용하는 것이다.

따라서 이는 사영추적 형태(projection pursuit type)의 통계량이라고 할 수 있다. $b_1(\boldsymbol{c})$를 $\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n$의 일변량 왜도의 제곱, 즉

$$b_1(\boldsymbol{c}) = \frac{n\left[\sum(y_i - \bar{y})^3\right]^2}{\left[\sum(y_i - \bar{y})^2\right]^3}$$

이라고 하자. 여기서 $y_i = \boldsymbol{c}'\boldsymbol{X}_i$이다. MA의 왜도는

$$b_{1,M} = \max_{\boldsymbol{c},\ \boldsymbol{c}\neq 0} b_1(\boldsymbol{c}) \tag{2.3}$$

로 정의되고, $b_{1,M} > K_1$이면 귀무가설 H_d를 기각한다. 또한 MA의 첨도는

$$b_{2,M}^2 = \max_{\boldsymbol{c},\ \boldsymbol{c}\neq 0} [b_2(\boldsymbol{c}) - 3]^2 \tag{2.4}$$

이고 $b_{2,M}^2 > K_2$이면 귀무가설을 기각한다. 여기서 $b_2(\boldsymbol{c})$는 $\boldsymbol{c}'\boldsymbol{X}_1, ..., \boldsymbol{c}'\boldsymbol{X}_n$의 일변량 첨도,

$$b_2(\boldsymbol{c}) = \frac{n\sum(y_i - \bar{y})^4}{\left[\sum(y_i - \bar{y})^2\right]^2}$$

이다. 이와 같은 사영추적형태의 통계량은 이론적인 연구가치는 충분하나 차원이 커질 경우 최대값을 갖는 벡터 $\boldsymbol{c}$를 찾아내는 것이 용이하지 않아 실제 사용에 제한을 받는다. Horswell & Looney(1992)에서도 $d > 2$일때는 실제로 MA의 왜도나 첨도는 이용하기 곤란하다는 사실을 언급하고 다변량 정규성 검정통계량의 비교연구에서 MA의 왜도나 첨도는 제외하였다.

통계량 $T = T(\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$이 H_d의 검정통계량일 때 모든 정칙행렬 $A \in R^{d\times d}$와 벡터 $\boldsymbol{b} \in \boldsymbol{R}^d$에 대해서

$$T(\boldsymbol{A}\boldsymbol{X}_1 + \boldsymbol{b}, ..., A\boldsymbol{X}_n + \boldsymbol{b}) = T(X_1, ..., X_n)$$

를 만족할 때 통계량 T는 affine invariance의 성질을 갖는다고 한다. $\boldsymbol{X}$가 정규분포를 따를 때 $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$도 역시 정규분포를 따르므로 이는 복합귀무가설 H_d의 검정을 위한 다변량 정규분포의 검정통계량의 바람직한 성질이라고 할 수 있다. 식(2.1), (2.2)의 Mardia의 왜도와 첨도, 식(2.3), (2.4)의 MA의 왜도와 첨도는 모두 affine invariance의 성질을 만족하고, 따라서 이들 통계량의 분포는 귀무가설하에서 $\boldsymbol{\mu}$와 $\boldsymbol{\Sigma}$에 의존하지 않는다. 또한 식(2.3), (2.4)의 $b_{1,M}$, $b_{2,M}^2$은 $||\boldsymbol{c}|| = 1$인 단위벡터에 대해서 최대를 고려하면 충분하다.

이 절에서는 MA의 왜도와 첨도의 근사통계량을 제안하고자 한다. 우선 MA의 왜도와 첨도의 계산방법을 살펴보자. $\boldsymbol{S}^*$를 $\boldsymbol{S}^{*\prime}\boldsymbol{S}\boldsymbol{S}^* = \boldsymbol{I}$를 만족하는 $n \times n$ 행렬이라고 하고, $\boldsymbol{Z}_j$를 척도화 잔차(scaled residuals)

$$\boldsymbol{Z}_j = \boldsymbol{S}^{*\prime}(\boldsymbol{X}_j - \bar{X}), \quad j = 1, ..., n \tag{2.5}$$

라고 하자. MA의 왜도는 affine invariance의 성질을 만족하므로 변환된 변수에 대해서 계산해도 무방하다. 이때 $\sum_j(\boldsymbol{c}'\boldsymbol{Z}_j)^2 = n$이므로

$$b_{1,M} = \max_{\boldsymbol{c},\ ||\boldsymbol{c}||=1} \frac{1}{n^2}\left[\sum(\boldsymbol{c}'\boldsymbol{Z}_j)^3\right]^2 \tag{2.6}$$

이 된다. $\sum(c'Z_j)^3$의 최대값을 얻기위하여 라그랑즈 승수법(Lagrange multiplier method)을 이용하면, $c'c = 1$이라는 제한조건과 함께

$$\sum(c'Z_j)^2 Z_j - \lambda c = 0 \tag{2.7}$$

를 얻는다. 식(2.7)의 양변에 c'을 곱하면 $\lambda = \sum(c'Z_j)^3$이 되고 따라서 $b_{1,M} = \lambda^2/n^2$이 된다. MA는 식(2.7)의 근사해를 얻기위하여 뉴튼-라프슨 방법(Newton-Raphson method)을 사용할 수 있다고 언급하고 있으나 이는 차원 d가 클 경우에는 현실성이 없어 보인다. 따라서 식(2.7)의 근사해, 즉

$$\frac{1}{n}\sum(c'Z_j)^2 Z_j = \frac{1}{n}\sum(c'Z_j)^3 c \tag{2.8}$$

의 근사해를 얻기위해서 이의 수렴값을 고려하면 합을 기대값으로 대치한 식

$$E\left[(c'Z)^2 Z\right] = E\left[(c'Z)^3 c\right] \tag{2.9}$$

을 얻게된다. 만일 $c = Z/\|Z\|$라면 식(2.9)의 좌우변은 모두 $E((Z'Z)^2 Z/\|Z\|^2)$으로 동일하다는 것을 쉽게 알 수 있다. 따라서 식(2.8)의 근사해로서 $c = Z_l/\|Z_l\|$, $l = 1, ..., n$을 고려할 수 있고 식(2.6)의 근사통계량으로

$$b^*_{1,M} = \max_{1\le l\le n}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\frac{Z_l'}{\|Z_l\|}Z_j\right)^3\right]^2 = \max_{1\le l\le n}\frac{1}{n^2}\frac{\left[\sum_{j=1}^{n}(Z_l'Z_j)^3\right]^2}{(Z_l'Z_l)^3}$$

을 제안한다. 다시말해서 단위 d-차원 구면(unit d-sphere) $S^{d-1} = \{c \in R^d : \|c\| = 1\}$의 모든 벡터 c에 대해서 최대를 고려하는 대신에 경험적 분포(empirical distribution)가 근사적으로 S^{d-1}에서의 균등분포(uniform distribution)를 따르는 표준화된 척도화 잔차 (normalized scaled residuals) $Z_l/\|Z_l\|$에 대해서 최대를 고려하자는 것이다. 이는 표준화된 척도화 잔차의 평균과 원점 사이의 거리를 고려하는 Koziol(1983)의 제안과도 공통점이 있다고 할 수 있다. 또한 Fang & Wong(1993)은 S^{d-1}에서 균등하게 분포되어 있는 단위벡터의 부분집합 $\{c_1, ..., c_N\}$을 선택하는 방법을 제안하였다. $X_1, ..., X_n$에 대해서 $b^*_{1,M}$는

$$b^*_{1,M} = \max_{1\le l\le n}\frac{1}{n^2}\frac{\left[\sum_{j=1}^{n}\left((X_l - \bar{X})'S^{-1}(X_j - \bar{X})\right)^3\right]^2}{\left[(X_l - \bar{X})'S^{-1}(X_l - \bar{X})\right]^3} \tag{2.10}$$

으로 표현된다.

MA의 첨도에 대해서도 같은 방법으로 근사통계량을 고려할 수 있다. 식(2.5)의 Z_j에 대해서

$$b^2_{2,M} = \max_{c,\ \|c\|=1}\left[\frac{1}{n}\sum(c'Z_j)^4 - 3\right]^2$$

이고 왜도와 마찬가지 방법으로

$$\sum(c'Z_j)^3 Z_j - \gamma c = 0$$

를 얻고 근사해로서 $\boldsymbol{c} = \boldsymbol{Z}_l/\|\boldsymbol{Z}_l\|, l = 1, ..., n$을 고려한다. 그러면 $b_{2,M}^2$의 근사통계량은

$$b_{2,M}^{2*} = \max_{1\le l\le n}\left[\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\boldsymbol{Z}_l'}{\|\boldsymbol{Z}_l\|}\boldsymbol{Z}_j\right)^4 - 3\right]^2 = \max_{1\le l\le n}\left[\frac{1}{n}\frac{\sum_{j=1}^{n}(\boldsymbol{Z}_l'\boldsymbol{Z}_j)^4}{(\boldsymbol{Z}_l'\boldsymbol{Z}_l)^2} - 3\right]^2$$

즉,

$$b_{2,M}^{2*} = \max_{1\le l\le n}\left[\frac{1}{n}\frac{\sum_{j=1}^{n}\left((\boldsymbol{X}_l - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})\right)^4}{\left[(\boldsymbol{X}_l - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_l - \bar{\boldsymbol{X}})\right]^2} - 3\right]^2 \tag{2.11}$$

이다.

식(2.10)의 $b_{1,M}^*$와 식(2.11)의 $b_{2,M}^{2*}$는 모두 일반화된 제곱 반지름(generalized version of the squared radii)

$$r_{ij} = \boldsymbol{Z}_i'\boldsymbol{Z}_j = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})'\boldsymbol{S}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})$$

의 함수이다. 따라서 두 통계량 모두 affine invariance의 성질을 갖는다(Henze(2002) Proposition 2.1 참조). 그리고 $b_{1,d}$와 $b_{1,M}^*$의 차이는 식(2.5)의 척도화 잔차를 정규화(normalized)했는가, 그리고 일반화된 제곱반지름의 평균을 취했는가 최대를 취했는가 하는 점이라고 할 수 있다.

$b_{1,M}$과 $b_{1,M}^*$, $b_{2,M}^2$과 $b_{2,M}^{2*}$의 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

$$D_1 = \frac{b_{1,M} - b_{1,M}^*}{b_{1,M}}, \quad D_2 = \frac{b_{2,M}^2 - b_{2,M}^{2*}}{b_{2,M}^2}$$

을 구하여 각 표본크기에서의 평균을 표 2.1과 표 2.2에 제시하였다. 여기서 $n = 10(10)50$은 10부터 50까지 10씩 증가함을 의미한다. 이로부터 표본크기가 커짐에 따라 상대오차가 현저하게 감소함을 볼 수 있고 따라서 $b_{1,M}^*, b_{2,M}^{2*}$는 각각 $b_{1,M}, b_{2,M}^2$의 합리적인 근사통계량이라고 할 수 있다.

표 2.1: 표본크기 $n = 10(10)50, 100$인 $N = 1000$개의 표본에서 계산된 상대오차 D_1의 평균

n	10	20	30	40	50	100
상대오차평균	0.05229	0.02684	0.01596	0.01028	0.005980	0.001972

3. 모의실험결과

이 절에서는 제안된 근사통계량의 근사백분위수와 몇개의 대립가설에서의 검정력을 알아보기 위한 모의실험을 행하였다.

표 2.2: 표본크기 $n = 10(10)50, 100$인 $N = 1000$개의 표본에서 계산된 상대오차 D_2의 평균

n	10	20	30	40	50	100
상대오차평균	0.07707	0.03486	0.01831	0.01232	0.008014	0.002566

식(2.3)과 식(2.4)의 MA의 왜도 $b_{1,M}$과 첨도 $b_{2,M}$의 귀무가설 H_d에서의 극한분포는 Machado(1983), de Wet, Venter & van Wyk(1979)에 의해 연구되었고 Baringhaus & Henze (1991)는 좀 더 일반적으로, 구형대칭(spherically symmetric)일 때 $b_{1,M}$과 $b_{2,M}$의 분포에 대해서 연구하였다. $b_{1,M}$과 $b_{2,M}$의 H_d에서의 분포는 다음과 같다. $(\boldsymbol{c}'\boldsymbol{c})^k$을 전개했을 때의 항의 수 $(k+d-1)!/k!(d-1)!$을 $N_{k,d}$라고 하고 $(\boldsymbol{c}'\boldsymbol{c})^k$를 전개했을 때의 계수의 제곱근을 $\boldsymbol{h}_{k,d}(\boldsymbol{c}) = (h_1(\boldsymbol{c}), ..., h_{N_{k,d}}(\boldsymbol{c}))'$이라고 하자. 예를 들어 $k = 3, d = 2$일 때 $N_{k,d} = 4$이고 $\boldsymbol{h}_{3,2}(\boldsymbol{c}) = (c_1^3, \sqrt{3}c_1^2c_2, \sqrt{3}c_1c_2^2, c_2^3)$이다. $\boldsymbol{Z}_{k,d} = (Z_1, ..., Z_{N_{k,d}})'$이고 $Z_1, ..., Z_{N_{k,d}}$는 독립이며 같은 분포를 따르는 $(i.i.d.)$ $N(0,1)$에서의 확률변수라고 할 때

$$\frac{n}{6}b_{1,M} \xrightarrow{d} \sup_{\boldsymbol{c},\ \|\boldsymbol{c}\|=1} \left(\boldsymbol{h}_{3,d}(\boldsymbol{c})'\boldsymbol{Z}_{3,d}\right)^2 := {W_{3,d}}^2 \tag{3.1}$$

$$\frac{n}{24}b_{2,M}^2 \xrightarrow{d} \sup_{\boldsymbol{c},\ \|\boldsymbol{c}\|=1} \left(\boldsymbol{h}_{4,d}(\boldsymbol{c})'\boldsymbol{Z}_{4,d}\right)^2$$

이 성립한다.

표 3.1에서는 $\sqrt{\frac{n}{6}}\sqrt{b_{1,M}}$의 분포를 계산하기 위하여 $N = 5000$개의 표본을 평균이 $\boldsymbol{0}$이고 분산-공분산 행렬 $\boldsymbol{\Sigma} = \boldsymbol{I}$인 정규분포로부터 추출하였다. 표본크기 $n = 10(10)50, 100$일 때 추출된 표본으로부터 통계량의 값을 계산하여 각 유의수준에서의 근사백분위수를 구하고 이를 Machado(1983)에 제시된 식(3.1)의 $W_{3,d} = \sqrt{W_{3,d}^2}$의 백분위수와 비교하였다. Marchado(1983)에 제시된 값은 표 3.1의 마지막 행$(n = \infty)$에 기록하였다. 비교결과 n이 약 50이상일 때 근사백분위수가 점근분포의 백분위수와 상당히 가까움을 볼 수 있다. 표 3.2에서는 $\sqrt{\frac{n}{6}}\sqrt{b_{1,M}^*}$에 해당하는 값을 같은 방법으로 구하였다. 표 3.1와 표 3.2의 분포를 비교해 보아 식(2.10)의 $b_{1,M}^*$도 역시 $b_{1,M}$과 같은 극한분포 (3.1)를 갖을 것이라고 예상하나 이에 대한 구체적이고 이론적인 연구는 좀 더 필요하다고 생각한다.

다음으로 $b_{1,M}^*$와 $b_{2,M}^{2*}$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 Mardia의 왜도 $b_{1,d}$와 첨도 $b_{2,d}$도 비교대상에 포함하였다. 그들은 (i) 주변분포가 서로 독립인 분포 (ii) 혼합정규분포(mixtures of normal distributions) 등을 고려하였다. 표 4.1, 4.2에서 $N(0,1)$, $C(0,1)$, $Logis(0,1)$, exp(1)은 각각 표준정규분포, 코쉬분포, 로지스틱분포, 지수분포를 나타낸다. χ_k^2과 t_k는 자유도가 k인 카이제곱분포와 t분포를 나타낸다. 또한 $\Gamma(a,b)$는 감마분포, $B(a,b)$는 베타분포, $LN(a,b)$는 대수정규분포를 나타낸다. 그리고 $F_1 * F_2$는 서로 독립인 주변분포 F_1과 F_2를 갖는 분포이며 F_1^2은 각각의 주변분포가 서로 독립인 F_1 분포임을 의미한다.

표 3.1: $\sqrt{\frac{n}{6}}\sqrt{b_{1,M}}$의 $100(1-\alpha)\%$ 근사백분위수 ($n = 10(10)50, 100$, $d = 2$)

n	α =0.25	0.10	0.05	0.025	0.001
10	1.5189	1.8960	2.1607	2.3726	2.5889
20	1.6893	2.1209	2.4833	2.7852	3.1372
30	1.7676	2.2021	2.5217	2.8525	3.3020
40	1.8128	2.3005	2.6072	2.9283	3.3257
50	1.8465	2.3346	2.6496	2.9790	3.3801
100	1.8849	2.3604	2.6532	2.9489	3.2550
∞	1.944	2.376	2.658	2.918	3.238

$NMIX_2(\kappa, \delta, \rho_1, \rho_2)$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 3.2: $\sqrt{\frac{n}{6}}\sqrt{b^*_{1,M}}$의 $100(1-\alpha)\%$ 근사백분위수 ($n = 10(10)50, 100$, $d = 2$)

n	α =0.25	0.10	0.05	0.025	0.001
10	1.4884	1.8820	2.1089	2.3062	2.5269
20	1.6930	2.1523	2.4662	2.7686	3.1363
30	1.7695	2.2339	2.5561	2.8794	3.3147
40	1.8123	2.2986	2.6565	2.9362	3.3137
50	1.8393	2.3095	2.6496	2.9768	3.3735
100	1.8841	2.3473	2.6524	2.9404	3.2561

표 4.1과 표 4.2의 검정력은 $N = 1000$개의 표본 중 유의한 표본의 백분위를 소수 첫째자리에서 반올림한 것이다. $b_{1,d}$, $b_{2,d}$의 경우 −는 Henze & Zirkler(1990)에 자료가 주어져 있지 않음을 의미한다. 표 4.1에서는 $b^*_{1,M}$, $b^{2*}_{2,M}$의 검정력을 MA의 왜도와 첨도 $b_{1,M}$, $b^2_{2,M}$의 검정력, Henze & Zirkler(1990)에 주어진 Mardia의 왜도와 첨도 $b_{1,d}$, $b_{2,d}$의 검정력과 비교하고 있다. 이로부터 제안된 근사통계량 $b^*_{1,M}$, $b^{2*}_{2,M}$와 MA의 통계량 $b_{1,M}$, $b^2_{2,M}$는 거의 같은 검정력을 가짐을 볼 수 있다. 또한 $b^*_{1,M}$, $b^{2*}_{2,M}$의 검정력은 Mardia의 왜도와 첨도 $b_{1,d}$, $b_{2,d}$의 검정력과도 매우 비슷한 양상을 보임을 알 수 있다. 왜도에 관해서는 고려된 대부분의 대립가설에서 $b_{1,d}$가 $b_{1,M}$이나 $b^*_{1,M}$보다 전반적으로 좋은 검정력을 보여주고 있다. 첨도에 관해서는 $b^2_{2,M}$이나 $b^{2*}_{2,M}$가 $b_{2,d}$보다 전반적으로 좋은 검정력을 나타낸다.

표 4.2에서는 $d = 5$일 때 $b_{1,M}$, $b_{2,M}^2$의 계산이 용이하지 않기 때문에 $b_{1,M}^*$, $b_{2,M}^{2*}$의 검정력을 $b_{1,d}$, $b_{2,d}$의 검정력과 비교하고 있다. 이 경우도 $d = 2$일 때와 유사한 경향을 발견할 수 있다. 그러나 $b_{1,M}^*$는 표본크기가 $n = 20$일 때 $d = 2, n = 20$일 경우와 비교하여 검정력의 감소가 두드러진다. 이는 차원과 표본크기의 비가 크기 때문에 나타나는 현상으로 짐작된다. 그러나 $b_{1,d}$의 경우는 이러한 경향을 발견할 수 없다. $b_{1,M}^*$도 $n = 50$일 때는 이러한 경향이 사라진다. $b_{2,M}^{2*}$가 $b_{2,d}$와 비교하여 우수한 검정력을 나타내는 경향도 표본크기와 자료의 차원이 커지면서 그 정도가 약해짐을 볼 수 있다.

$b_{2,d}$의 경우, 주변분포의 꼬리가 정규분포보다 얇은 베타분포에서 다른 통계량들은 거의 귀무가설을 기각하지 못하는데 비해 $b_{2,d}$는 상대적으로 우수한 검정력을 보여준다.

4. 결론 및 토의

본 논문에서는 Roy의 합교원리(union-intersection principle)을 이용하여 왜도와 첨도를 다변량으로 확장한 MA 통계량을 근사적으로 구하는 방법에 대해서 고려하였다. 제안된 근사통계량은 d-차원 구면 S^{d-1}의 모든 벡터를 고려하는 대신에 경험적 분포가 근사적으로 S^{d-1}에서의 균등분포를 따르는 확률변수를 이용하였고 affine invariance의 성질을 유지한다.

Mardia의 왜도, 첨도와 MA의 왜도, 첨도는 같은 일변량 통계량을 각각 다른 방법으로 다변량으로 확장한 것으로 비정규성의 규명에 있어서 어떤 차이가 있는지에 대한 연구가 필요하다. 이에 본 논문에서 제안한 MA의 왜도와 첨도의 근사통계량이 도움이 될 것으로 생각한다. 이들의 형태는 척도화 잔차의 함수이다. MA의 왜도의 근사통계량의 경우, Mardia의 왜도와 상당히 비슷한 형태를 갖는다. 3절의 모의실험 결과로 볼 때 근사통계량의 분모, 즉 척도화 잔차의 정규화가 검정력의 향상에 전혀 도움을 주지 못하는 것으로 생각되나 이에 대해서는 좀 더 이론적인 규명이 필요하다.

참고문헌

김남현 (2004). 다변량 정규성검정을 위한 근사 Shapiro-Wilk 통계량의 일반화. <응용통계연구>, **17**, 35-47.

Baringhaus, L. and Henze, N. (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika,* **35**, 339-348.

Baringhaus, L. and Henze, N. (1991). Limit distributions for measures of multivariate skewness and kurtosis based on projections. *Journal of Multivariate Analysis,* **38**, 51-69.

Baringhaus, L. and Henze, N. (1992). Limit distributions for Mardia's measure of multivariate skewness. *The Annals of Statistics,* **20**, 1889-1902.

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

de Wet, T., Venter, J. H. and van Wyk, J. W. J. (1979). The null distributions of some test criteria of multivariate normality. *South African Statistical Journal,* **13**, 153-176.

Fang, K. T. and Wang, Y. (1993). *Number-theoretic methods in statistics.* Monographs on statistics and applied probability. Chapman and Hall, London.

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

Henze, N. (2002). Invariant tests for multivariate normality : A critical review. *Statistical Papers,* **43**, 467-506.

Henze, N. and Wagner, T. (1997). A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis,* **62**, 1-23.

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.

Horswell, R. L. and Looney, S. W. (1992). A comparison of tests for multivariate normality that are based on measures of multivariate skewness and kurtosis. *Journal of Statistical Computation and Simulation,* **42**, 21-38.

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

Koziol, J. A. (1983). On assessing multivariate normality. *Journal of the Royal Statistical Society, Series B,* **45**, 358-361.

Machado, S. G. (1983). Two statistics for testing for multivariate normality. *Biometrika,* **70**, 713-718.

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

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

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

Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Applied Statistics,* **24**, 163-171.

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

Pearson, E. S., D'Agostino, R. B. and Bowman, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika,* **64**, 231-246.

Romeu, J. L. and Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis,* **46**, 309-334.

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

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

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

Zhu, L., Fang, K. T. and Bhatti, M. I. (1997). On estimated projection pursuit Cramér-von Mises statistics. *Journal of Multivariate Analysis,* **63**, 1-14.

Zhu, L. X., Wong, H. L. and Fang, K. T. (1995). A test for multivariate normality based on sample entropy and projection pursuit. *Journal of Statistical Planning and Inference,* **45**, 373-385.

[2004년 1월 접수, 2004년 5월 채택]

표 4.1: 각 분포에서 $b_{1,d}$, $b_{1,M}$, $b^*_{1,M}$, $b_{2,d}$, $b^2_{2,M}$ $b^{2*}_{2,M}$ 통계량의 검정력 비교 (유의수준 $\alpha = 0.05$, $n = 20,\ 50$, $d = 2$)

	$n = 20$						$n = 50$					
대립가설	$b_{1,d}$	$b_{1,M}$	$b^*_{1,M}$	$b_{2,d}$	$b^2_{2,M}$	$b^{2*}_{2,M}$	$b_{1,d}$	$b_{1,M}$	$b^*_{1,M}$	$b_{2,d}$	$b^2_{2,M}$	$b^{2*}_{2,M}$
$N(0,1)^2$	6	4	5	5	4	5	5	6	5	5	4	4
$\exp(1)^2$	80	73	71	46	51	55	100	100	100	82	84	86
$LN(0,.5)^2$	60	54	53	34	40	41	97	95	95	68	72	73
$C(0,1)^2$	93	91	92	96	95	96	–	99	99	–	100	100
$\Gamma(0.5,1)^2$	–	92	91	–	72	74	–	100	100	–	97	96
$\Gamma(5,1)^2$	25	26	24	14	16	17	68	60	60	27	31	29
$(\chi^2_5)^2$	43	40	39	22	28	29	92	88	85	48	53	53
$(\chi^2_{15})^2$	19	17	17	10	12	15	49	45	43	18	24	25
$(t_2)^2$	67	64	64	68	70	71	91	90	87	97	96	98
$(t_5)^2$	25	24	28	22	28	27	46	42	39	54	54	56
$B(1,1)^2$	0	0	1	36	0	0	0	0	0	91	0	0
$B(1,2)^2$	7	6	7	12	3	3	23	16	17	25	1	1
$B(2,2)^2$	1	1	2	12	1	1	0	0	0	44	0	0
$Logis(0,1)^2$	17	12	16	14	16	15	24	23	22	27	32	33
$N(0,1) * \exp(1)$	47	49	47	26	30	31	93	93	93	51	59	62
$N(0,1) * \chi^2_5$	23	28	24	12	19	19	61	65	61	23	32	32
$N(0,1) * t_5$	17	17	16	13	17	18	23	24	27	24	33	33
$N(0,1) * B(1,1)$	2	3	2	10	2	3	2	2	3	29	2	2
$NMIX_2(.5,2,0,0)$	3	3	3	9	2	3	2	2	2	16	3	3
$NMIX_2(.5,4,0,0)$	3	3	3	22	2	2	2	2	2	48	2	2
$NMIX_2(.5,2,.9,0)$	32	22	22	11	19	16	80	52	51	18	37	33
$NMIX_2(.5,.5,.9,0)$	22	17	14	14	19	18	30	27	26	23	37	35
$NMIX_2(.5,.5,.9,-.9)$	42	33	37	27	32	35	65	54	53	46	63	66

표 4.2: 각 분포에서 $b_{1,d}$, $b^*_{1,M}$, $b_{2,d}$, $b^{2*}_{2,M}$ 통계량의 검정력 비교 (유의수준 $\alpha = 0.05$, $n = 20,\ 50$, $d = 5$)

	$n = 20$				$n = 50$			
대립가설	$b_{1,d}$	$b^*_{1,M}$	$b_{2,d}$	$b^{2*}_{2,M}$	$b_{1,d}$	$b^*_{1,M}$	$b_{2,d}$	$b^{2*}_{2,M}$
$N(0,1)^5$	5	4	4	6	5	6	5	5
$\exp(1)^5$	82	59	55	57	100	99	95	91
$LN(0,.5)^5$	63	44	42	44	100	94	87	82
$C(0,1)^5$	99	98	99	100	-	100	-	100
$\Gamma(0.5,1)^5$	98	83	85	81	-	100	-	99
$\Gamma(5,1)^5$	20	15	11	19	70	50	35	34
$(\chi^2_5)^5$	40	30	21	28	96	80	61	57
$(\chi^2_{15})^5$	15	11	9	14	49	39	22	27
$(t_2)^5$	86	78	85	81	99	98	100	100
$(t_5)^5$	34	28	24	30	64	51	68	60
$B(1,1)^5$	0	1	28	1	0	0	90	0
$B(1,2)^5$	3	3	7	4	10	2	20	1
$B(2,2)^5$	1	1	13	2	0	1	47	0
$Logis(0,1)^5$	15	12	11	16	35	30	36	30
$N(0,1)^4 * \exp(1)$	19	21	11	18	62	59	26	40
$N(0,1)^4 * \chi^2_5$	10	11	7	11	29	32	13	19
$N(0,1)^4 * t_5$	9	9	7	10	18	19	15	22
$N(0,1)^4 * B(1,1)$	-	4	-	4	2	4	8	3

Remarks on the Use of Multivariate Skewness and Kurtosis for Testing Multivariate Normality *

Namhyun Kim [1)]

ABSTRACT

Malkovich & Afifi (1973) generalized the univariate skewness and kurtosis to test a hypothesis of multivariate normality by use of the union-intersection principle. However these statistics are hard to compute for high dimensions. We propose the approximate statistics to them, which are practical for a high dimensional data set. We also compare the proposed statistics to Mardia(1970)'s multivariate skewness and kurtosis by a Monte Carlo study.

Keywords: Goodness of fit; Multivariate normality; Skewness; Kurtosis.

* 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, Department of Science, Hongik University, 72-1 Sangsu-dong, Mapo-gu, Seoul, 121-791, Korea.
E-mail: nhkim@hongik.ac.kr