# 다변량 정규성검정을 위한 근사 SHAPIRO-WILK 통계량의 일반화 *

김남현[1)]

요 약

Fattorini(1986)의 통계량은 Shapiro와 Wilk의 일변량 정규분포를 위한 검정통계량을 다변량으로 확장한 것이다. 본 논문에서는 Kim과 Bickel(2003)에서 제안한 이변량 정규분포를 위한 검정통계량을 Fattorini(1986)의 방법을 이용하여 이변량 이상인 경우에도 실제적으로 사용가능하도록 일반화하였다. 제안된 통계량은 Fattorini(1986) 통계량의 근사통계량으로 생각할 수 있으며 표본의 크기가 클 때도 사용가능하다.

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 $d$는 $d \geq 1$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. 다변량 정규성 검정에 대한 일반적인 방법에 대해서는 Mardia(1980), Thode(2002, Chapter 9) 그리고 D'Agostino and Stephens(1986, section 9.7) 등을 참고로 한다.

Malkovich와 Afifi(1973), Fattorini(1986)은 Shapiro와 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. Kim과 Bickel(2003)에서는 Shapiro와 Wilk(1965)의 검정통계량과 밀접한 관련이 있고, 같은 극한분포를 갖는 de Wet과 Vener(1972)의 일변량 정규분포의 검정통계량을 Malkovich와 Afifi(1973)에서와 마찬가지로 Roy의 union-intersection원리를 이용하여 이변량으로 일반화하였다. 또한 제안된 통계량의 귀무가설에서의 극한분포

* 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

를 가우스 과정(Gaussian process)의 적분의 형태로 표현하고 모의실험을 통하여 다른 통계량과의 검정력을 비교하였다.

Kim과 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 Kim과 Bickel(2003)의 통계량을 수정, 보완하고자 한다. 그 결과 제안된 통계량은 $d \geq 2$인 임의의 $d-$변량에서 사용가능하게 된다.

## 2. Malkovich와 Afifi(MA)와 Fattorini(FA)의 검정

일변량 정규분포의 검정을 위한 Shapiro-Wilk의 통계량(Shapiro와 Wilk(1965)) $W$는

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

이다. 여기서 $Z_{(1)}, \ldots, Z_{(n)}$은 일변량 확률표본 $Z_1, \ldots, Z_n$의 순서통계량, $\bar{Z}$는 표본평균이고 $a_j$는 Shapiro와 Wilk(1965)에 주어진 상수이다. Malkovich와 Afifi(1973)이 제안한 방법은 적절한 상수 $K_w$에 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(Shapiro와 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. $l$은 $\{1, \ldots, n\}$에서의 임의의 정수이므로 $n$개의 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

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

을 만족하는 $c^{(m)}$을 택하였다. 따라서 MA 통계량은

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

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 $n$개의 가능한 해인 $\boldsymbol{c}^{(1)},...,\boldsymbol{c}^{(n)}$에 대해서조차 $W(\boldsymbol{c})$을 최소화하지 못한다는데 착안하여 Fattorini(1986)은

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

을 제안하였다. 여기서 $U_{(j)}$는

$$U_j = (\boldsymbol{X}_l - \bar{\boldsymbol{X}})'\mathbf{A}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}}), \quad j = 1,...,n, \tag{2.5}$$

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 Shapiro와 Wilk(1965)에 주어져 있으므로, MA와 FA는 $n \le 50$일 때 사용가능하다.

## 3. 제안된 검정통계량

일변량 정규분포의 검정을 위한 de Wet과 Venter(1973)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 $s^2$은 표본분산 $s^2 = n^{-1}\sum(Z_i-\bar{Z})^2$이고 $H_i = \Phi^{-1}(\frac{i}{n+1})$, $\Phi^{-1}$는 표준정규분포 $N_1(0,1)$의 분포함수 $\Phi$의 역함수이다.

de Wet과 Venter의 $L_n$-통계량은 일변량 정규성 검정을 위한 Shapiro와 Wilk(1965)의 $W$-통계량, Shapiro와 Francia(1972)의 $W'$-통계량과 밀접한 관련이 있다. 사실상 $L_n$-통계량은 $W$-통계량의 간단한 형태로 생각될 수 있고 (D'Agostino and Stephens(1986, 5.10절), de Wet and Venter(1972)), 세 통계량은 모두 같은 근사분포를 갖는다는 것이 증명되었다(Leslie, Stephens과 Fotopolous(1986)). 이들 통계량의 극한분포에 대해서는 de Wet and Venter(1972), Csörgő(1983, 7장), del Barrio, Cuesta, Matrán and Rodíguez(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 $r_n$의 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 $r_n^2$은 식 (2.1)의 $W$-통계량과 계수 $a_j$를 제외하고 같은 형태로 표현된다. 따라서 Shapiro와 Wilk(1965)의 Lemma 3과 유사한 다음의 보조정리를 얻을 수 있다.

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

증명: Shapiro와 Wilk(1965)의 Lemma 3의 증명과 같은 방법을 적용한다. □

Kim and Bickel(2002)에서는 $d = 2$, $\boldsymbol{X} = (X_1, X_2)'$일때 복합귀무가설 $H_2$를 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 $i$번째 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 $(c_1, c_2) = (\cos\theta, \sin\theta)$, $0 \le \theta \le \pi$로 가정해도 무방하다. 즉, 이변량 분포일 때는 $P_n$-통계량을 계산할 때 단일변수 $\theta$에서 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 $P_n$은 이변량뿐만 아니라 $d > 2$인 다변량에서도 같은 방법으로 정의될 수 있다. 그러나 $d > 2$인 다변량에서는 $d = 2$인 경우와 달리, $P_n$의 계산이 실제적으로 용이하지 않다. 이 절에서는 이러한 $P_n$의 단점을 해결하기 위해서 $P_n$의 근사통계량을 Fattorini(1986)의 방법을 이용하여 제안하고자 한다.

식(3.3)에 의해서

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

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

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

라고 하면 $P_n$이 클때 귀무가설 $H_d$를 기각하는 검정은 $R_n^2$이 작을 때 $H_d$를 기각하는 검정과 동일하다.

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

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

$$\begin{aligned} P_n^* &= \min_{1\le l\le n} L_n(c^{(l)}) \\ &= \min_{1\le l\le n} \sum_{i=1}^n \left( \frac{U_{(i)} - \bar{U}}{(\frac{1}{n}(\boldsymbol{X}_l - \bar{\boldsymbol{X}})'\boldsymbol{A}^{-1}(\boldsymbol{X}_l - \bar{\boldsymbol{X}}))^{1/2}} - H_i \right)^2 \end{aligned} \tag{3.5}$$

을 얻을 수 있다. 여기서 $U_{(j)}$는 식(2.5)의 $U_j$의 순서통계량이다.

식(3.4)의 $P_n$-통계량과 식(3.5)의 $P_n^*$-통계량의 근사정도를 보기 위하여 모의실험을 행하였다. 이변량 정규분포 $N_2(\mathbf{0}, \boldsymbol{I})$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 3.1에 제시하였다. 이로부터 표본크기가 커짐에 따라 상대오차가 현저하게 감소함을 볼 수 있고 따라서 $P_n^*$는 $P_n$의 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 |

## 참고문헌

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

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

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

[4] 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.

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

[6] 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.

[7] Kim, N. and Bickel, P. J. (2003). The limit distribution of a test statistic for bivariate normality. *Statistica Sinica, 13, (to be appeared)*

[8] 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.

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

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

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

[12] 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.

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

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