응용통계연구 제 15권 2호, 2002년, pp. 223-232

가능도 함수를 기초로 한 다변량 정규성 검정 *

여인권[1)]

요 약

이 논문에서는 비선형 변환과 가능도 함수를 이용하여 다변량 자료의 정규성을 검정하는 방법에 대해 알아본다. 사용된 변환은 변환모수에 따라 여러 가지 형태를 가지는 변환족을 구성하는데 이 변환모수를 검정하여 자료의 정규성을 검정한다. 모수의 검정은 점수함수(score function)을 기초로 이루어지며 표본크기가 적은 경우에도 검정통계량의 분포를 유도하기 위한 모수적 붓스트랩 검정방법이 사용된다. 모의실험 결과 기존의 방법과 검정력을 비교하여 제안된 방법이 검정력이 높은 것으로 나타났다.

주요용어: 멱변환, 붓스트랩 가설검정, 점수검정.

1. 서론

일반적인 통계분석에서 많이 사용되고 있는 모수적 기법들은 대부분 자료들이 특정한 분포를 따른다는 가정을 한 후, 그 분포의 모수에 대해 추론을 하는 것으로 결과를 도출한다. 이러한 모수적 분석은 자료들이 얼마나 가정한 분포와 일치하는가에 따라 사용된 분석방법의 타당성이 결정되기 때문에 모수적 방법을 사용하기 위해서는 자료와 설정한 가정의 적합성을 검진할 필요가 있다.

과거 수십 년간 수많은 통계적 분석방법들이 개발되어 왔지만 정규분포를 기반으로 개발된 분석 방법들이 실제 자료 분석에 있어 가장 많이 사용되고 있다. 정규분포의 경우, 평균과 분산에 대한 분석만으로 모든 결과를 도출할 수 있으며 모수에 대한 추정량과 검정통계량의 통계적 특성이 많이 알려져 있어 통계적 추론이 다른 분포를 가정했을 때보다 쉬어진다는 장점이 있다. 그러나 한 가지 주의해야 할 점은 평균과 분산은 이상점에 민감하게 반응하기 때문에 이들 값을 기본으로 이루어지는 분석 또한 이상점에 의해 잘못된 결과를 유도할 수 있다는 단점이 있다. 자료가 꼬리부분이 한쪽으로 길게 뻗어져 있는 경우나 양쪽 꼬리부분이 두터운 경우라면, 그 만큼 이상점이 발생할 가능성이 높아지고 이에 따라 분석에 영향을 미칠 가능성이 높아진다. 이러한 이유 때문에 특별히 로버스트한 방법을 이용하여 자료를 분석하지 않는 한 자료의 정규성을 확인할 필요가 있다. 이 절에서는 지금까지 소개된 대표적 정규성 검정방법들을 특징에 따라 분류하고 그 내용에 대해 간단히 알아본다.

* 이 논문은 2002년도 전북대학교의 지원 연구비에 의해 연구되었음

1) 전라북도 전주시 덕진구 덕진동 1가 전북대학교 자연과학대학 수학통계정보과학부 조교수
E-mail: inkwon@moak.chonbuk.ac.kr

정규성 검정은 왜도와 첨도를 고려한 Pearson에 의해 시작되었다. 평균을 중심으로 k차 적률을 $\mu_k = E(X-\mu)^k$라고 하고 표본적률을 $m_k = \sum_{i=1}^n (X_i - \bar{X})^k/n$라고 하자. 정규분포의 경우, 왜도계수는 $\sqrt{\beta_1} = \mu_3/\mu_2^{3/2} = 0$이고 첨도계수는 $\beta_2 = \mu_4/\mu_2^2 = 3$이기 때문에 $\sqrt{\beta_1} \neq 0$이거나 $\beta_2 \neq 3$이라는 것은 정규분포를 따르지 않는다는 것을 의미한다. 그러므로 표본분산을 S^2라고 할 때, 표본 왜도계수 $\sqrt{b_1} = m_3/S^3$와 표본 첨도계수 $b_2 = m_4/S^4$가 각각 0과 3에서 얼마나 떨어져 있는가를 확인하여 정규성의 판단 근거로 사용될 수 있다. 이 두 통계량을 혼합한 통계량이 D'Agostino와 Pearson(1973), Cox와 Hinkley(1974) 그리고 Bowman과 Shenton(1975) 등에 의해 제안되었다.

Kolmogrov와 Smirnov는 경험적 분포함수 $F_n(\cdot)$을 기초로 한 검정통계량을 소개하였다. 함수 $\Phi(\cdot)$을 표준정규분포함수라고 하고 $X_{(i)}$ 를 i번째 순서통계량이라고 하면 Kolmogrov-Smirnov 검정통계량은 다음과 같이 정의된다.

$$K = \max_{-\infty<x<\infty} |F_n(x) - \Phi((x-\hat{\mu})/\hat{\sigma})| = \max(D^+, D^-),$$

여기서 $D^+ = \max_{1\le i\le n} \{i/n - \Phi((X_{(i)}-\hat{\mu})/\hat{\sigma})\}$이고 $D^- = \max_{1\le i\le n} \{\Phi((X_{(i)}-\hat{\mu})/\hat{\sigma}) - (i-1)/n\}$를 나타낸다. 또한 $Y_i = \Phi((X_{(i)} - \hat{\mu})/\hat{\sigma})$를 이용한 많은 적합도 검정방법이 소개되었는데 Cramer-von Mises 검정통계량 $W^2 = 1/(12n) + \sum_{i=1}^n [Y_i - (2i-1)/(2n)]^2$, Anderson-Darling 검정통계량 $A^2 = -n - \sum_{i=1}^n [(2i-1)(\log(Y_i) + \log(1-Y_{n+1-i}))/n]$, 그리고 Kuiper 검정 통계량 $V = D^+ + D^-$등이 대표적인 방법들이다. 이들 통계량들은 정규분포뿐만 아니라 일반적인 연속형 확률분포에서도 사용할 수 있다.

Shapiro와 Wilk(1965)은 정규분포를 가정하여 Gauss-Markov 정리에 의해 구해진 분산의 최량선형비편향추정량(BLUE)와 S^2를 비교한 W 통계량을 소개하였다. 이 통계량은 D'Agostino(1971), Shapiro와 Francia(1972), Weisberg와 Bingham(1975), Puri와 Rao(1976)등에 의해 수정 보완되었다. 이 중 SAS와 같은 통계프로그램에서 많이 사용되고 있는 방법은 표준정규분포로부터 추출된 i번째 순서통계량의 기대값인 표준정규점수(normal score) s_i와 정규점수의 상관관계를 이용한

$$W = \frac{\{\sum_{i=1}^n (s_i - \bar{s})(X_{(i)} - \bar{X})\}^2}{\sum_{i=1}^n (s_i - \bar{s})^2 \sum_{i=1}^n (X_{(i)} - \bar{X})^2},$$

가 사용되고 있다.

위에서 언급한 방법들은 대부분 단일변량에 한정되어 개발되었고 비록 다변량으로 확장시킨 통계량들이 제안되었지만 검정통계량의 분포를 유도하기 어려울 뿐만 아니라 표본의 수가 적은 경우에는 유도된 분포와 차이가 많은 문제점이 있다. 다변량 정규성 검정에 관련된 참고문헌은 http://science.ntu.ac.uk/msor/mjb/mvnbib.html를 참고하기 바란다. 이 논문에서는 단일변량 뿐만 아니라 다변량에서도 쉽게 사용할 수 있는 정규성 검정방법에 대해 알아본다. 제안된 방법은 정규성뿐만 아니라 일반적인 연속형 확률분포에도 쉽게 적용할 수 있다는 장점을 가진다. 이변량 분포를 중심으로 한 모의실험을 통해 기존에 사용되고 있는 방법과 제안된 방법의 검정력을 비교해 본다.

2. 정규성 검정을 위한 제안된 방법론

통계분석에서 분포의 가정이 만족되지 않을 때, 일반적으로 많이 사용되는 방법 중에 하나는 설정한 가정에 더욱 만족하도록 자료를 변환시킨 후 분석하는 것이다. Box와 Cox(1964)는 모수 값에 따라에 다양한 형태를 가지는 멱변환족(family of power transformations)을 이용하여 자료가 정규분포를 따르지 않을 때 자료를 정규분포에 가깝게 변환시킨 후 통계분석을 하는 방법을 제안하였다. 이 때 사용된 변환이 Box-Cox 변환인데 이 변환은 변환 모수에 대해 연속하고 단조성을 가지기 때문에 가능도 함수를 사용하여 모수에 대한 추론을 할 수 있다. 참고로 Box-Cox 변환의 경우 관측값이 모두 양수이여야 한다는 제약 조건이 있기 때문에 이 논문에서 제안한 방법에서 가정한 변환으로 활용하는데 무리가 있을 수 있다.

이 논문에서는 아래의 두 조건을 만족하는 임의의 비선형 변환 $h(\lambda, x)$을 중심으로 일반적인 이론을 도출한다.

I. 변환 $h(\lambda, x)$는 변환모수 λ에 대해 2차 미분이 가능하고
II. 어떤 특정 값 $\lambda = \lambda_0$에 대해, $h(\lambda_0, x) = x$가 된다.

John과 Draper(1980)와 Yeo와 Johnson(2000)는 이러한 성질을 만족하는 변환족을 소개하였는데 3절에서 이들 변환의 특성에 대해 알아보고 어떻게 제안된 방법에 사용되는지를 알아본다. 여기서 주의할 점은 조건 II에서 $\lambda = \lambda_0$일 때에는 자료를 변환할 필요가 없다는 것을 의미한다.

임의의 p차원 연속형 다변량 분포로부터 추출된 확률벡터를 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$이라고 하자. 변환 h는 비선형이기 때문에 자료의 형태가 동일하더라도 위치와 척도에 따라 λ에 대한 추론 결과가 다르게 나올 수 있다. 즉, 위치불변성(location invariance)과 척도불변성(scale invariance)을 만족하지 않을 수 있다. 이러한 문제점을 해결하기 위해 추출된 각각의 확률변수는 표본평균이 0이고 표본분산이 1로 표준화시킨 값을 사용한다.

변환모수 $\boldsymbol{\lambda}$에 대한 추론은 Box와 Cox(1964)가 선형모형에서 가정했던 것처럼 임의의 변환모수벡터 $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$에 대해 변환된 확률벡터 $h(\boldsymbol{\lambda}, \boldsymbol{X}_1), \ldots, h(\boldsymbol{\lambda}, \boldsymbol{X}_n)$는 각각 평균이 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$이고 공분산행렬이

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_{pp} \end{pmatrix}$$

인 다변량 정규분포를 따른다고 가정 하에서 이루어진다. 여기서 변환된 각각의 확률벡터는 $h(\boldsymbol{\lambda}, \boldsymbol{X}_i) = (h_1(\lambda_1, X_{i1}), \ldots, h_p(\lambda_p, X_{ip}))$로 정의된다. 표기상의 편의를 위해 먼저 변환모수를 포함한 모수벡터를 $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$로 분해하는데 여기서 $\boldsymbol{\theta}_1 = \boldsymbol{\lambda}$를 나타내고 $\boldsymbol{\sigma}$를 공분산행렬의 원소로 이루어진 $p(p+1)/2$차원 벡터라고 할 때 $\boldsymbol{\theta}_2 = (\boldsymbol{\mu}, \boldsymbol{\sigma})$를 의미한다. 이 가정

하에서 로그 가능도 함수는

$$l_n(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}(h(\boldsymbol{\lambda},\boldsymbol{x}_i)-\boldsymbol{\mu})\,\Sigma^{-1}\,(h(\boldsymbol{\lambda},\boldsymbol{x}_i)-\boldsymbol{\mu})^T + \sum_{i=1}^{p}\sum_{j=1}^{n}\log|(J(\lambda_i,x_{ij}))|, \quad (2.1)$$

가 되는데 여기서 det는 행렬식을 의미하고 $J(\lambda_i,x_{ij})$는 x_{ij}로부터 $h_i(\lambda_i,x_{ij})$로의 변환의 야코비안 행렬을 나타낸다.

조건 II에 의하면 만약 자료가 이미 다변량 정규분포를 따르고 있다면 $\boldsymbol{\lambda}$의 추정량 $\hat{\boldsymbol{\lambda}}$는 $\boldsymbol{\lambda}_0$ 근처에 있을 것이다. 그러므로 추정량 $\hat{\boldsymbol{\lambda}}$이 얼마나 $\boldsymbol{\lambda}_0$ 가까이에 있는가를 확인하여 자료가 정규분포를 따르는지 아닌지를 확인할 수 있을 것이다. 이 논문의 기본 아이디어는 자료의 정규성을 확인하기 위해 가설 $H_0:\boldsymbol{\lambda}=\boldsymbol{\lambda}_0$의 사실여부를 검정하는 것이다. 이와 같은 상황에서 가설을 검정하기 위한 일반적인 방법으로는 Wald 검정, 가능도비 검정과 점수 검정(score test)을 고려할 수 있다. 이 논문에서는 계산상으로 $\boldsymbol{\lambda}$에 대한 추정량을 구하지 않고도 검정통계량을 계산할 수 있는 편의성을 고려하여 점수 검정을 사용할 것이다.

귀무가설 $H_0:\boldsymbol{\lambda}=\boldsymbol{\lambda}_0$하에서 최대 가능도 추정량을 $\hat{\boldsymbol{\theta}}_0=(\boldsymbol{\lambda}_0,\hat{\boldsymbol{\mu}}_0,\hat{\boldsymbol{\sigma}}_0)$라고 하면 앞 절에서 언급한 것과 같이 각각의 확률변수는 표준화되었기 때문에 $\hat{\boldsymbol{\mu}}_0=\bar{\boldsymbol{X}}=\mathbf{0}$이고 $\hat{\boldsymbol{\sigma}}_0$는 행렬 $n^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^T(\boldsymbol{X}_i-\bar{\boldsymbol{X}})$에서 $\boldsymbol{\sigma}$와 대응되는 값들의 벡터가 된다. 점수함수(score function)는 로그 가능도 함수 (2.1)를 각각의 모수에 대해 미분한 값으로

$$S_{ij_i}(\boldsymbol{\theta}) = \frac{\partial l_n(\boldsymbol{\theta}|\boldsymbol{X})}{\partial\theta_{ij_i}}, \qquad i=1,2,\ j_1=1,\ldots,p,\ j_2=1,\ldots,p(p+3)/2$$

로 정의되는데 가설을 검정에 사용될 검정통계량 $\boldsymbol{S}_1(\hat{\boldsymbol{\theta}}_0)=(S_{11}(\hat{\boldsymbol{\theta}}_0),\ldots,S_{1p}(\hat{\boldsymbol{\theta}}_0))$가 $\mathbf{0}$에서 멀리 떨어질수록 자료의 정규성을 의심할 수 있다. 귀무가설 하에서 $\boldsymbol{S}_1(\hat{\boldsymbol{\theta}}_0)$는 평균이 $\mathbf{0}$이고 공분산 행렬이 $\Sigma(\hat{\boldsymbol{\theta}}_0)=\mathbf{I}_{11}(\hat{\boldsymbol{\theta}}_0)-\mathbf{I}_{12}(\hat{\boldsymbol{\theta}}_0)\mathbf{I}_{22}^{-1}(\hat{\boldsymbol{\theta}}_0)\mathbf{I}_{12}(\hat{\boldsymbol{\theta}}_0)^T$인 정규분포에 근사한다. 여기서

$$\begin{aligned}\mathbf{I}_{11}(\hat{\boldsymbol{\theta}}_0) &= E[S_1(\hat{\boldsymbol{\theta}}_0)S_1(\hat{\boldsymbol{\theta}}_0)^T],\\ \mathbf{I}_{12}(\hat{\boldsymbol{\theta}}_0) &= E[S_1(\hat{\boldsymbol{\theta}}_0)S_2(\hat{\boldsymbol{\theta}}_0)^T],\\ \mathbf{I}_{22}(\hat{\boldsymbol{\theta}}_0) &= E[S_2(\hat{\boldsymbol{\theta}}_0)S_2(\hat{\boldsymbol{\theta}}_0)^T]\end{aligned}$$

로 정의되며 $\hat{\boldsymbol{\theta}}_0$에서의 Fisher 정보행렬(Fisher information matrix)을 분할한 벡터와 행렬을 나타낸다. 그러나 $\Sigma(\hat{\boldsymbol{\theta}}_0)$는 간명하게 계산되지 않기 때문에 관측정보행렬(observed information matrix)을 이용하여 추정해야 한다. Fisher정보행렬의 값과 대응하는 관측정보행렬의 값을 $\mathbf{J}_{11}(\hat{\boldsymbol{\theta}}_0)$, $\mathbf{J}_{12}(\hat{\boldsymbol{\theta}}_0)$, 그리고 $\mathbf{J}_{22}(\hat{\boldsymbol{\theta}}_0)$라고 하면 Patefield(1977)의 결과를 이용하여 분산의 추정량을

$$\hat{\Sigma}(\hat{\boldsymbol{\theta}}_0)=\mathbf{J}_{11}(\hat{\boldsymbol{\theta}}_0)-\mathbf{J}_{12}(\hat{\boldsymbol{\theta}}_0)\mathbf{J}_{22}^{-1}(\hat{\boldsymbol{\theta}}_0)\mathbf{J}_{12}(\hat{\boldsymbol{\theta}}_0)^T=\left(-\frac{\partial^2 l_n(\boldsymbol{\theta}|\boldsymbol{x})}{\partial\lambda_i\partial\lambda_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_0}\right),\quad i,j=1,\ldots,p$$

로 구할 수 있다. 결론적으로 장애모수(nuisance parameter)인 μ와 σ가 수반된 상황에서 가설 $H_0: \lambda = \lambda_0$를 검정하기 위한 검정통계량은 근사적으로

$$T(\hat{\theta}_0) = S_1(\hat{\theta}_0)\hat{\Sigma}(\hat{\theta}_0)^{-1}S_1(\hat{\theta}_0)^T \sim \chi_p^2 \tag{2.2}$$

가 된다. 여기서 단일변량의 경우 검정통계량 $S_1(\hat{\theta}_0)/\sqrt{\hat{\Sigma}(\hat{\theta}_0)}$를 사용하여 λ에 대해 단측검정도 가능하다. 만약 검정 결과가 자료들이 정규분포를 따르지 않는다고 판단되어 정규성을 가정한 분석방법을 사용하기 어려운 경우에는 사용된 변환을 사용하여 $\hat{\lambda}$을 구한 다음 자료를 변환하여 정규분포에 가깝게 만든 후 분석할 수 있다. 또한 주의해야 할 점은 로그가능도 함수 (2.1)를 다른 분포의 로그가능도 함수로 대체하여 사용해도 같은 결과를 얻기 때문에 정칙조건하의 일반적인 연속형 확률분포에 모두 적용시킬 수 있다는 것이다.

검정 통계량 $T(\hat{\theta}_0)$의 정확한 분포는 표본의 크기가 상당히 크기 않는 한 카이제곱분포와 다른 경우가 많다. 이런 경우를 대비해서 실제 활용에서는 붓스트랩 검정(Bootstrap test)을 추천하고 싶다. 이 붓스트랩 방법은 이 논문에서 제안한 방법뿐만 아니라 앞 절에서 언급한 모든 통계량에 대해 활용할 수 있어 다음 절에 나올 모의실험에서 붓스트랩 검정을 이용하여 검정 방법들을 비교할 것이다.

붓스트랩 검정의 기본 과정은 다음과 같다. 검정통계량 $T(\theta)$의 분포함수를 $F(\theta, s) = P(T(\theta) \leq s)$라고 하자. 함수 $F(\hat{\theta}_0, s)$는 단조 증가하며 연속성이 있기 때문에 유의수준 $\alpha \times 100\%$의 임계값 C에 대한 붓스트랩 추정량은 $\hat{C} = F^{-1}(\hat{\theta}_0, 1-\alpha)$으로 구할 수 있다. 여기서 $\hat{C}$는 다음과 같은 붓스트랩 샘플링을 통해 근사할 수 있다:

1. 표준화된 관측값으로부터 표본상관행렬을 계산한다.
2. 과정 1에서 계산된 상관행렬을 공분산 행렬로 가지는 다변량 정규분포로부터 독립적인 n개의 p 차원 정규 붓스트랩 표본 $\boldsymbol{X}_b^* = (\boldsymbol{X}_{1b}^*, \ldots, \boldsymbol{X}_{nb}^*)$, $b = 1, \ldots, B$,를 발생시킨다.
3. 각각의 붓스트랩 표본 $\boldsymbol{X}_b^*$를 다시 표준화시킨 후 표준화된 표본을 이용하여 최대가능도 추정량 $\hat{\theta}_{0b}^*$와 $T(\hat{\theta}_{0b}^*|\boldsymbol{X}_b^*)$를 계산한다.
4. 붓스트랩 검정통계량 $T(\hat{\theta}_0^*|\boldsymbol{X}^*)$들로부터 $\hat{C}^* = (1-\alpha)$분위수를 계산하여 기각역 $[\hat{C}^*, \infty)$를 구하고 표본으로부터 계산된 $T(\hat{\theta}_0|\boldsymbol{x})$가 기각역에 있는지를 확인하여 자료의 정규성을 검정한다.

정칙조건하에서 최대가능도추정량 $\hat{\theta}_0$은 일치성을 만족하기 때문에 Beran(1986)에 의하면 이 붓스트랩 검정은 점근적으로 옳은 수준(correct level)을 가진다.

3. 변환선택과 모의실험

자료의 정규성을 결정할 때 많은 경우 분포의 중간 형태보다는 꼬리부분의 형태가 영향을 더 주는 경향이 있다. 그러므로 자료의 정규성을 적절하게 확인하려면 꼬리부분의 특성을 잘 반영하는 변환을 선택해야 한다. 제안된 방법에서 사용될 적절한 변환을 선택하기

위해서 van Zwet(1964)의 볼록-오목(convex-concave) 변환의 개념을 먼저 간단히 알아본다. Van Zwet에 의하면 전체적으로 볼록한 변환은 분포의 오른쪽을 당기고 왼쪽을 밀어내는 효과가 있기 때문에 분포가 양의 왜도인 경우 이 변환을 사용하면 왜도를 줄여주고 반대로 오목한 변환은 음의 왜도를 줄여주는 효과가 있다고 한다. 또한 오른쪽은 볼록하고 왼쪽은 오목한 변환은 분포의 양쪽을 중앙으로 당겨주는 효과가 있어 꼬리부분이 두꺼운 경우 첨도를 줄여주는 효과가 있다.

이 절에서는 대립분포가 한쪽으로 기울어져 있는 경우와 대칭이지만 양쪽 꼬리부분이 길게 뻗어져 있는 형태를 고려하여 아래의 두 변환족을 선택하여 모의실험을 실시했다.

$$h_1(\lambda, x) = \begin{cases} \left\{(x+1)^\lambda - 1\right\}/\lambda & \lambda \neq 0, x \geq 0, \\ \log(x+1) & \lambda = 0, x \geq 0, \\ -\left\{(-x+1)^{2-\lambda} - 1\right\}/(2-\lambda) & \lambda \neq 2, x < 0, \\ -\log(-x+1) & \lambda = 2, x < 0. \end{cases} \tag{3.1}$$

$$h_2(\lambda, x) = \begin{cases} \left\{(x+1)^\lambda - 1\right\}/\lambda & \lambda \neq 0, x \geq 0, \\ \log(x+1) & \lambda = 0, x \geq 0, \\ -\left\{(-x+1)^\lambda - 1\right\}/\lambda & \lambda \neq 0, x < 0, \\ -\log(-x+1) & \lambda = 0, x < 0. \end{cases} \tag{3.2}$$

변환 (3.1)은 Yeo와 Johnson(2000)에 의해 소개된 변환족으로 $\lambda < 1$일 때 전체적으로 볼록하고 $\lambda > 1$일 때 전체적으로 오목한 형태를 가지기 때문에 왜도된 대립분포(skewed alternative distribution)인 경우에 사용될 수 있다. 우리가 흔히 접하는 자료들은 양의 왜도를 가지는 경우인데, 단일변량의 경우 $\lambda < 1$인지를 확인하는 단측 검정을 사용해도 된다. 변환 (3.2)는 John과 Draper(1980)에 의해 소개된 율변환족(modulus transformation)으로 분포가 대칭일 때 정규분포에 가깝게 만들기 위해 사용되었으며 대칭인 대립분포(symmetric alternative distribution)에서 사용될 수 있다. 변환 모수가 $\lambda < 1$이면 변환의 오른쪽은 볼록하고 왼쪽은 오목해져 이상점이 빈번히 발생할 수 있는 두터운 꼬리부분을 가졌는지를 확인하는데 사용될 수 있다. 주의해야 할 점은 두 변환 모두 앞 절에서 언급한 조건 I과 II를 만족하는데 $\lambda = 1$일 때 항등함수가 되는 것을 볼 수 있다.

이변량에서 제안된 방법이 기존의 방법에 비해 얼마나 검정력이 높은지를 확인하기 위해 앞에서 언급한 붓스트랩 검정을 이용한 모의실험을 통해 비교해 보았다. 검정력의 객관성을 고려하여 모든 검정법에 대해 같은 붓스트랩 표본을 이용하여 결과를 얻었다. 양의 왜도가 있는 대립분포인 감마분포와 대칭이지만 꼬리부분이 두터운 t 분포를 대상으로 변환 (3.1)을 이용한 점수통계량 T_1과 (3.2)의 T_2와 Mardia(1970)의 다변량 왜도계수 $\sqrt{b_1}$와 첨도계수 b_2를 각각 비교하였다. Malkovich와 Afifi(1973)과 Giorgi와 Fattorini(1976)의 검정력 실험결과에 의하면 Mardia의 계수는 Malkovich와 Afifi(1973)이 Roy의 union-intersection 원리를 이용하여 다변량에 일반화시킨 Shapiro-Wilk의 W 통계량과 함께 검정력이 다른 방법들 보다 높은 것으로 나타났으며 새롭게 개발될 검정방법은 이들 방법과 비교해야 할 것이라고 하였다. Mardia의 왜도와 첨도는 프로그래밍이 용이하여 SAS Macro로 작성되어 사용되고 있는데 http://ftp.sas.com/techsup/download/stat/multnorm.sas에서 다운로드 할 수 있다. 이 논문에서는 제안된 방법과 Mardia의 방법의 검정력을 비교하였다.

표 3.1: 대립가설하에서의 모의실험에 의한 검정력.

H_1	n	T_1	T_2	$\sqrt{b_1}$	b_2
감마(5)	20	0.709	0.302	0.418	0.203
	40	0.956	0.466	0.797	0.363
	80	1.000	0.656	1.000	0.582
감마(3)	20	0.728	0.316	0.457	0.200
	40	0.978	0.437	0.854	0.337
	80	1.000	0.627	0.997	0.594
지수	20	0.875	0.473	0.683	0.382
	40	0.997	0.669	0.971	0.646
	80	1.000	0.877	1.000	0.854
t(10)	20	0.193	0.201	0.065	0.050
	40	0.232	0.296	0.114	0.084
	80	0.350	0.469	0.159	0.128
t(5)	20	0.685	0.810	0.584	0.616
	40	0.857	0.984	0.834	0.905
	80	0.951	1.000	0.946	0.996
코쉬	20	0.931	0.982	0.896	0.950
	40	0.985	1.000	0.983	0.999
	80	0.996	1.000	0.996	1.000

코쉬를 제외한 이변량 난수의 상관관계가 0.5가 되도록 하였고 표본의 크기는 $n = 20$, 40, 80를 사용하였고 붓스트랩 표본을 200번 발생시켜 5% 유의수준으로 기각역을 구했으며 이 과정을 1000번 반복하여 검정력을 계산하였다. 다변량 난수를 발생하기 위한 알고리즘은 Johnson과 Kotz(1972)를 참고하였다. 표 1에서 괄호안의 값은 감마분포의 경우 형태모수의 값을 나타내고 t 분포의 경우 자유도를 나타낸다.

표 3.1은 붓스트랩 모의실험 결과를 보여준다. 이 표를 요약 정리하면 다음과 같은 결론에 도달한다.

1. 기울어진 대립분포인 경우, T_1과 $\sqrt{b_1}$이 높은 검정력을 제공하고 있으며 모든 결과에서 T_1이 월등히 높은 것을 볼 수 있다.

2. 꼬리부분이 두터운 대립분포인 경우, T_2과 b_2이 높은 검정력을 제공하고 있으며 T_2가 전반적으로 높은 것을 볼 수 있다.

이러한 모의실험 결과를 볼 때 기존에 사용되고 있는 방법들보다 제안된 방법이 비정규성을 더욱 잘 판명해 준다고 할 수 있다.

실제 자료 분석에 있어 제안된 방법을 사용하기 위해서는 먼저 단일변량에 대해 h_1과

h_2의 중 적절한 변환을 선택하고 전체적으로 정규성을 검정하는 단계를 거쳐야 한다. 이와 같은 과정을 자동으로 실행하여 검정해주는 프로그램을 저자의 홈페이지에서 다운로드 받을 수 있다. 또한 T_1과 T_2를 결합시킨 형태의 통계량을 이용하거나 $h_1(\lambda_1, h_2(\lambda_2, x))$나 $h_2(\lambda_2, h_1(\lambda_1, x))$와 같이 두 변환이 결합된 상황에서 $\lambda_1 = \lambda_2 = 1$를 검정하는 방법도 생각해 볼 수 있다.

참고문헌

[1] Beran, R. (1986). Simulated Power Functions. *Annals of Statistics* **14**, 151-173.

[2] Bowman, K. O. and Shenton, B. R. (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika* **62**, 243-250.

[3] Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, B* **26**, 211-252.

[4] Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall. London.

[5] D'Agostino, R. B. (1971). An Omnibus Test of Normality for Moderate and Large Sample Sizes. *Biometrika* **58**, 341-348.

[6] D'Agostino, R. B. and Pearson, E. S. (1973). Tests for departures from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika* **60**, 613-622.

[7] Giorgi, G. M. and Fattorini, L. (1976). An Empirical study of some tests for multivariate normality. *Quaderni dell'Instituto di Statistica* **20**, 1-8.

[8] John, J.A. and Draper, N.R. (1980). An Alternative Family of Transformations. *Applied Statistics* **29**, 190-197.

[9] Johnson, N. L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons, Inc. New York.

[10] Malkovich, J. F. and Afifi, A. A. (1973). On Tests for Multivariate Normality. *Journal of the American Statistical Association* **68**, 176-179.

[11] Mardia, K. V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. *Biometrika* **57**, 519-530.

[12] Patefield, W. M. (1977). On the Maximized Likelihood Function. *Sankhyā B* **39**, 92-96.

[13] Puri, M. L. and Rao, C. R. (1976). Augmenting Shapiro-Wilk Test for Normality Contributions to Applied Statistics. *Birkhauser (Grossohans)*, Berlin, 129-139.

[14] 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.

[15] Shapiro, S. S. and Wilk, M.B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591-611.

[16] van Zwet, W. R. (1964). *Convex Transformations of Random Variables.* Amsterdam: Mathematisch Centrum.

[17] Weisberg, S. and Bingham, C. (1975). An Approximate Analysis of Variance Test for Non-normality Suitable for Machine Calculation. *Technometrics* **17**, 133-134.

[18] Yeo, I.K. and Johnson, R.A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, **87**, 954-959.

[2001년 2월 접수, 2002년 3월 채택]

A Test of the Multivariate Normality Based on Likelihood Functions*

In-Kwon Yeo [1)]

ABSTRACT

The present paper develops a test of the multivariate normality based on non-linear transformations and the likelihood function. For checking the normality, we test the shape parameter which indexes the family of transformations. A score test and a parametric bootstrap test are used to evaluate the discrepancy between the data and a multivariate normal distribution. In order to compare the performance of our test with the existing tests, a simulation study was carried out for several situations where nuisance parameters have to be estimated. The results showed that the proposed method is superior to the existing methods.

Keywords: Power transformation, Bootstrap test, Score test

* This paper was supported (in part) by research funds of Chonbuk National University

1) Assistant Processor, Division of Mathematics and Statistical Informatics, Chonbuk National University.
E-mail: inkwon@moak.chonbuk.ac.kr