응용통계연구 제 14권 2호, 2001년, pp. 223-231

다변량 정규성과 이상치 검정을 위한 통계 시스템 개발 *

최용석 [1] 김종건 [2] 강명래[3]

요 약

다변량분석 기법을 사용하기 위해서는 자료가 정규성(normality)가정을 만족해야한다. 본 연구에서는 GUI환경에서 일변량 및 다변량자료의 정규성검정, 이상치제거 및 변수변환을 하는 시스템을 Visual Basic 언어로서 구축하여 사용자들이 보다 편리하게 사용할 수 있음을 소개 하고자 한다.

주요용어: 정규성, 이상치, GUI, Visual Basic.

1. 서론

실생활에서 우리가 접하는 자료들은 대부분 하나 이상의 변수들로 이루어져 있는 다변량 자료이고 다변량적 분석기법(인자분석, 대응분석, 판별분석 등)을 요구하는 자료들이 대부분이다. 이 다변량적 분석기법을 사용하기 위해서는 자료가 정규성의 가정을 만족해야 하며, 가정을 만족한다면 우리는 정규분포가 가지는 유용한 성질들을 이용하여 분석을 편리하게 할 수 있다. 그 첫째 이유가 정규분포는 이상적인 분포(Bona fide population)이고 둘째가 다변량 통계량의 샘플링분포는 모집단의 분포와 관계없이 관측치의 개수가 많을 경우 중심극한 정리(central limit theorem)에 의해서 정규분포를 따른다고 가정할 수 있다는 것이다(Johnson and Wichern, 1998, p. 157). 그러나 만약 자료가 정규성 가정을 만족하지 않는 경우에는 비모수적 방법이나 기타 분석방법을 통해서 분석하여야 하며 이런 분석방법들은 쉽지 않으리라 생각된다.

일변량 자료인 경우 정규성 검정을 SAS, SPSS, MINITAB 등 통계패키지에서 제공하는 정규확률그림(Q-Q plot), 왜도 및 첨도 나 샤피로-윌크(Shapiro-Wilk)검정통계량 등을 이용하여 쉽게 할 수 있다. 일반적으로 이전의 통계패키지에서는 다변량 자료의 정규성 검정에 관한 어떠한 정보도 제공받을 수 없었다. 최근 SAS/ETS v. 8.1에서는 왜도, 첨도, 헨즈-지클러(Henze-Zirkler)에 의한 검정방법이 소개 되고 있다. 그러나 통계전문가가 아니면 다변량 자료의 정규성을 검정하기 어렵고, 통계전문가 또한 복잡한 다변량 정규성 검정의 가정을 경시하거나 이 절차를 생략하고 분석하는 경우가 많아 통계적 오류를 가져오는 경우가 많다. 따라서 본 연구에서는 다음과 같은 두 가지 목적을 고려할 수 있다.

* 본시스템의 사용을 원하는 경우 http://home.pusan.ac.kr/~ yschoi 를 접속하면 된다.

1) (609-735) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 부교수, 컴퓨터및 정보통신연구소, 연구원 E-mail: yschoi@hyowon.pusan.ac.kr.

2) (617-737) 부산시 북구 구포3동, 부산정보대학 정보통신계열, 부교수
E-mail: jgkim@pitc.ac.kr.

3) (609-735) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 석사과정,
E-mail: banny99@hanmail.net.

첫째, 다변량 자료의 정규성 검정은 기존의 통계패키지에서 분석 프로시저를 지원하지 않기 때문에 분석의 어려움이 많았다. 이 문제를 해결하기 위해 통계적 시스템을 구축하였으며, GUI(graphic user interface)환경에서 통계적 지식이 없는 사람도 누구나 손쉽게 일변량 자료뿐만 아니라 다변량 자료에 대한 정규성 검정을 할 수 있게 하였다.

둘째, 정규성 가정을 만족하지 않는 자료에 대해서는 적합한 변수변환을 이용하여 자료가 정규분포에 접근할 수 있도록 하였으며, 더불어 여러 가지 이상치를 검정하는 방법을 통하여 이상치를 찾아내고, 제거함으로써 자료가 정규성에 접근할 수 있도록 하였다.

참고로 통계학 이론을 바탕으로 한 통계시스템 개발의 국내사례로는 허문열(1995), 유종영 외 2인(1997) , 서혜선 외 2인(1999), 최용석 · 현기홍(2000) 등이 있다.

2. 시스템의 구성

본 시스템은 비주얼 베이직 언어를 사용하여 GUI환경에서 누구나 쉽게 클릭만으로 사용할 수 있도록 하는데 중점을 두고 구축하였으며, 전체 시스템은 풀다운 형식의 메뉴(pull down menu)로 되어있다. 또한 시스템의 알고리즘에서 사용하고 있는 통계분포에 대한 적분 값을 비주얼 베이직 언어에서 지원하지 않기 때문에 수치 해석적인 방법을 통하여 근사적으로 구하여 사용하였다(Shammas, 1996, pp. 231-245).

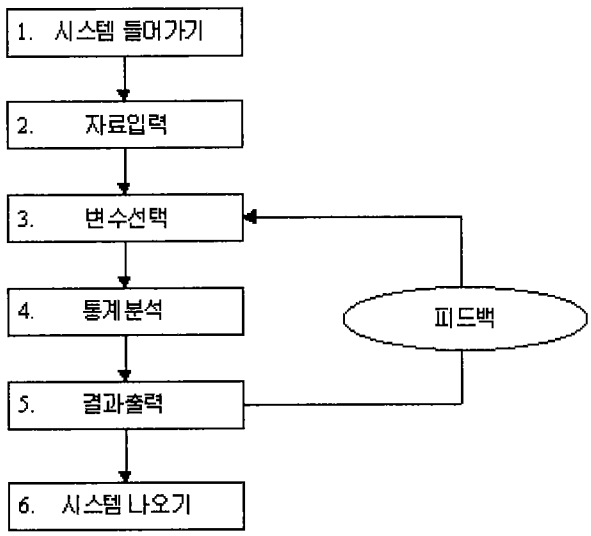

그림 2.1: 시스템의 자료처리순서

2.1. 시스템 들어가기

시스템을 설치하면 아래 그림 2.2 시스템의 로그화면이 나타나며, 로그화면을 클릭하

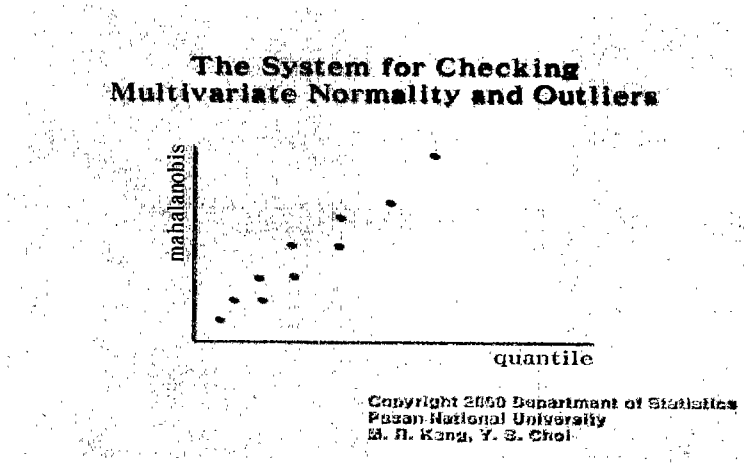

그림 2.2: 시스템의 로그화면

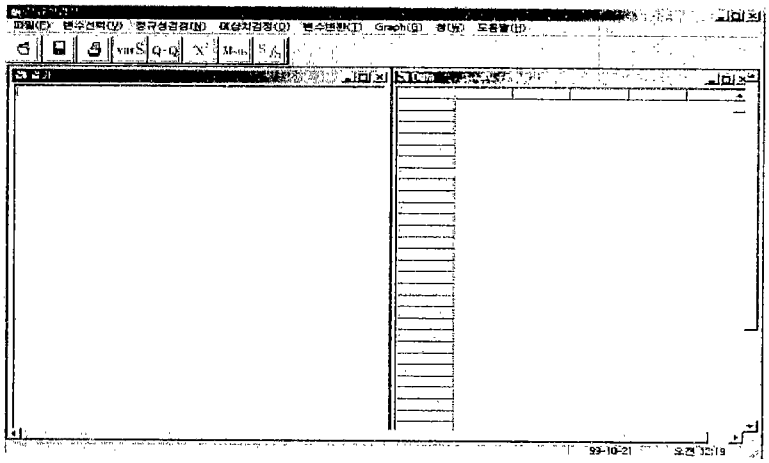

그림 2.3: 시스템의 주화면

면 그림 2.3 시스템의 주 화면이 나타난다. 만약 주 화면이 나타나지 않으면 설치에 이상이 있는 것으로 다시 설치하여야 한다.

그림 2.3 시스템의 주화면 위쪽에는 파일, 변수선택, 정규성, 이상치, 변수변환, 창, 도움말등 7가지의 풀다운 형식의 메뉴가 있고, 사용의 편리함을 위하여 여러 가지 기본 아이콘을 제공하고 있다. 그리고 자료입력을 위해 스프레드시트(spreadsheet)형식의 입력 창과, 자료출력과 그래프 지원을 위한 결과 창이 있다.

2.2. 자료입력 및 변수선택

그림 2.3에의 파일→입력 메뉴나 파일입력의 단축아이콘을 사용하여 분석에 사용할 자료(dat, txt형식의 파일)를 불러올 수가 있다. 불러온 자료가 입력 창에 입력이 되면, 변수선택 창에서 분석에 사용할 변수선택을 하고 분석을 할 수 있다.

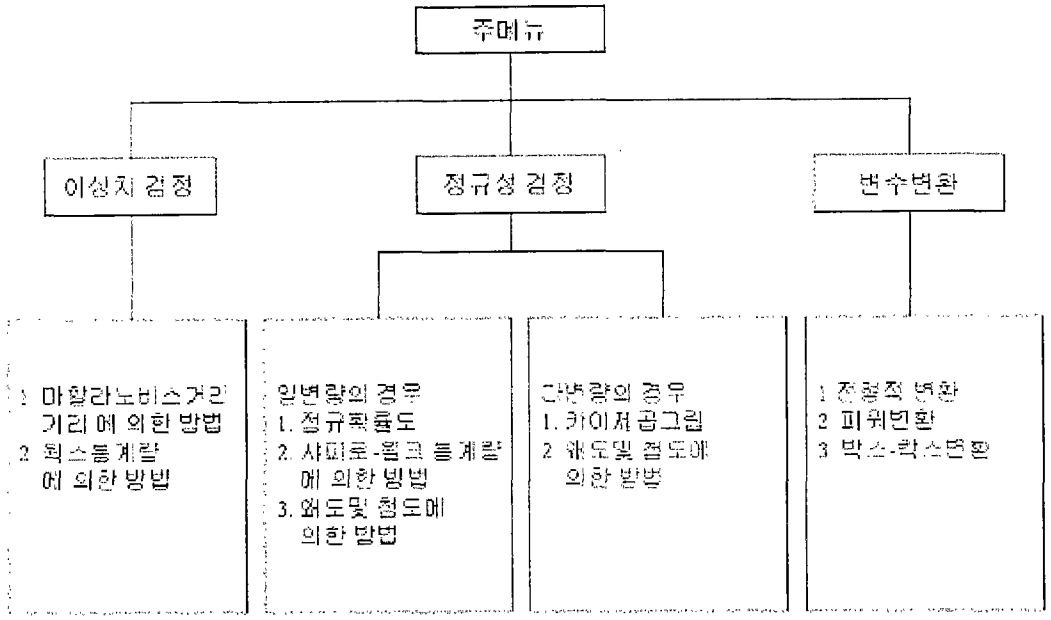

그림 2.4: 시스템의 분석 구성도

2.3. 통계분석

시스템에서 제공하는 통계처리 방법은 정규성 검정, 이상치 검정, 변수변환의 3가지이다. 정규성 검정은 일변량 자료와 다변량 자료의 분석알고리즘이 다르므로 이를 구분하여 분석하도록 하였다. 그리고 이상치 제거는 일변량 및 다변량 자료 모두 같은 분석을 실행할 수 있으며, 변수변환은 일변량에 대한 변수변환 방법을 반복적으로 적용하였다. 그림 2.4 시스템의 분석구성도에서는 시스템이 지원하고 있는 통계처리 프로시저를 항목별로 나열하고 있다. 그리고 이와 관련된 통계량, 수식 그리고 통계적 이론은 Jobson(1992, pp. 148-152), Johnson and Wichern (1998, pp. 194-214), Rencher(1995, pp.107-118)을 참고하였다.

2.4. 결과출력 및 시스템 나오기

통계분석을 하고 나면 결과 창에 기초통계량 및 분석결과가 나타난다. Q-Q 플롯, 카이

표 3.1: 네 가지 강도 측정자료

board	X1	X2	X3	X4	board	X1	X2	X3	X4
1	1889	1651	1561	1778	16	1954	2149	1180	1281
2	2403	2048	2087	2197	17	1325	1170	1002	1176
3	2119	1700	1815	2222	18	1419	1371	1252	1308
4	1645	1627	1110	1533	19	1828	1634	1602	1755
5	1976	1916	1614	1883	20	1725	1594	1313	1646
6	1712	1712	1439	1546	21	2276	2189	1547	2111
7	1943	1685	1271	1671	22	1899	1614	1422	1477
8	2104	1920	1717	1874	23	1633	1513	1290	1516
9	2983	2794	2412	2581	24	2061	1867	1646	2037
10	1745	1600	1384	1508	25	1856	1493	1356	1533
11	1710	1591	1518	1667	26	1729	1412	1238	1469
12	2046	1907	1627	1898	27	2168	1896	1701	1834
13	1840	1841	1595	1741	28	1655	1675	1414	1597
14	1867	1685	1493	1678	29	2326	2301	2065	2234
15	1859	1649	1389	1714	30	1490	1382	1214	1284

제곱 그림과 박스-칵스 변환의 분석을 사용한다면 그림 창에 그림이 나타난다. 만약 이상치 제거나 변수변환등의 자료의 변형이 있었다면 입력 창에 변형된 자료가 생성됨을 확인할 수 있다. 그리고 필요한 결과에 대해서 그림 2.3의 파일→저장 및 파일→인쇄 메뉴를 통하여 결과를 출력하고 파일→종료 메뉴로부터 시스템을 빠져 나올 수 있다.

3. 자료를 이용한 시스템의 실행 예

이 장에서는 표 3.1의 실제 자료를 사용하여 본 시스템의 실행과정을 살펴본다. 이 자료는 30개의 판자(board)의 강도를 4가지 다른 측도(X1, X2, X3, X4)를 통하여 얻어진 자료이다. X1은 충격파(shock wave)를 판자에 보냈을 때 얻어진 측도 이며, X2는 판자가 흔들리고 있는 동안에 얻어진 측도이며, X3, X4는 정적인 상태에서 얻어진 측도이다(Johnson and Wichern, 1998, p. 198).

시스템이 제공하는 여러 가지 분석 방법중 몇 가지 분석만 사용하여 그 사용방법 및 결과를 알아보도록 하자.

먼저 시스템을 구동시켜 주 화면이 나타나면 그림 2.3의 파일→입력 메뉴로부터 위의 예제 데이터를 입력 받을 수 있다. 정상적으로 자료를 입력받았다면 시스템의 입력 창에 자료입력창가 같이 자료가 입력되어있는 것을 확인할 수 있을 것이다. 다변량 자료의 정규성 검정을 시행해보자. 변수선택 메뉴로부터 4개의 변수 X1, X2, X3, X4를 선택하여 정규

성→카이제곱 그림 메뉴를 선택하면 카이제곱 그림의 결과창과 그림 3.1 카이제곱 그림을 볼 수 있을 것이다. 그리고 그림 2.3의 정규성→왜도 및 첨 도 메뉴를 선택하면 그림 3.2 왜도 및 첨도를 이용한 정규성 검정의 결과를 볼 수 있다.

다음으로 이상치 검정은 이상치→마할라노비스 거리 메뉴를 선택하면 유의수준 입력창이 뜰 것이다. 적절한 유의수준을 입력하고 확인버튼을 누르면 그림 3.3 마할라노비스 거리에 의한 이상치 검정의 결과 창을 볼 수 있다.

마지막으로 변수변환의 경우에는 그림 2.3의 변수변환→박스-칵스 메뉴를 선택하면 박스-칵스변환의 결과 창과 박스-칵스 그림을 볼 수 있다. 그리고 결과로부터 적절한 변환 값을 찾아 변환하여 입력 창에 변환된 값이 입력되어있는 것을 볼 수 있다.

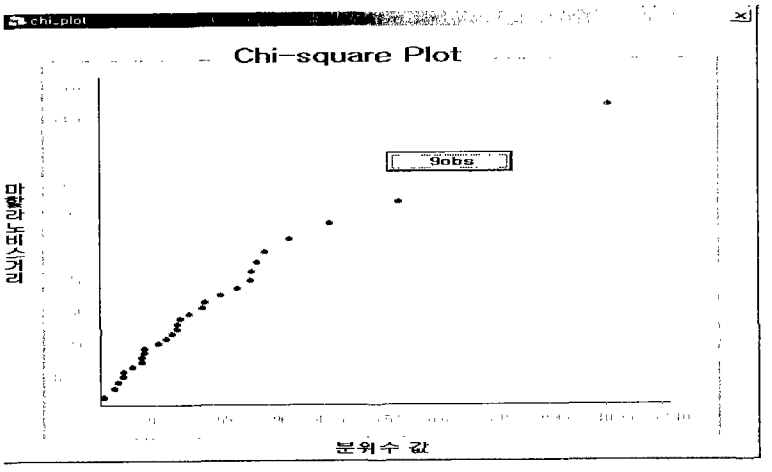

그림 3.1: 카이제곱 그림

4. 결론

다변량 자료를 분석하는데 앞서 가장 필요한 절차인 정규성 가정을 검정을 하기 위해 대부분 통계분석자들은 SAS/IML을 사용하여 힘들게 검정을 하거나 이 가정을 생략하여 통계적 오류를 일으키는 경우가 많았다. 이를 극복하기 위해서 본 시스템을 구축함으로 GUI 환경 하에서 통계적 지식이 없는 일반인들도 다변량 자료의 정규성 검정을 할 수 있으며 정규성을 만족하지 않는 자료에 대해서 이상치 제거나 변수변환을 통하여 정규성에 접근하도록 자료를 변형시킬 수 있도록 하였다. 본시스템의 성능을 좀더 향상시키기 위해서 몇 가지 발전 방향을 제시한다면, 시스템이 지원하는 이상치 검정방법은 하나의 이상치를 반복적으로 검정하는 방법만을 선택하여 제공하였다. 따라서 이상치 검정에 있어 하나씩 이상치를 반복적으로 검정할 때, 각각은 이상치 이지만 다른 이상치의 효과에 의하여 이상치로 검출되지 않는 가면효과(masking effect)가 생길 수도 있다. 본 연구에서는 이런 이 시스

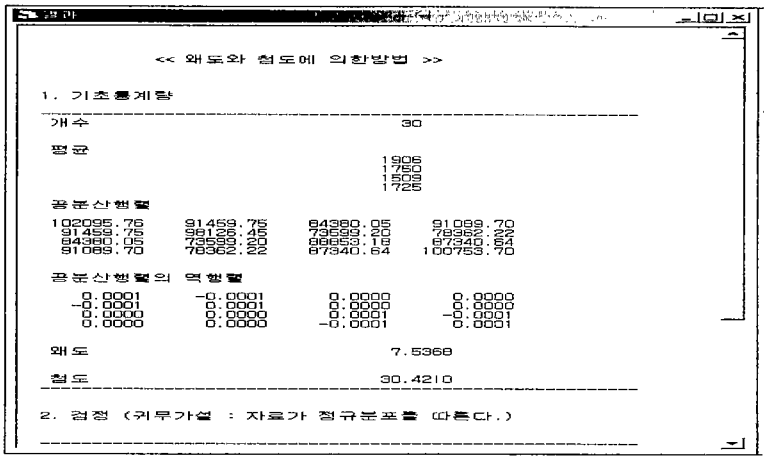

그림 4.1: 왜도 및 첨도에 의한 검정의 결과

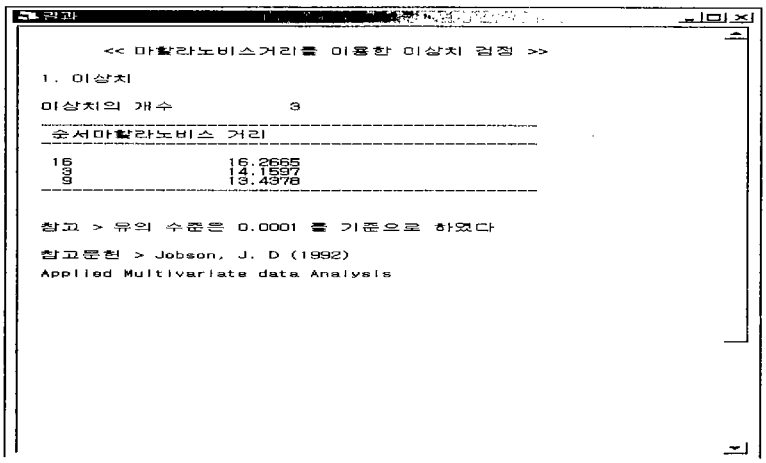

그림 4.2: 마할라노비스 거리에 의한 이상치 검정의 결과

템에서는 가면효과를 찾을 수 있는 다중 이상치(multiple outliers)검정을 지원하는 프로시저를 구축하지 못하였다. 이런 점들을 좀더 보완한 다면 보다 나은 시스템이 될 수 있을 것이다.

참고문헌

[1] 서혜선, 김미경, 허명회 (1999). SAS AF/SCL로 구현한 다변량 수량화 시스템, <한국분류학회>, 제 3권, 1-11.

[2] 유종영, 안기수, 허문열 (1997). 동적 그래픽스에 의한 회귀진단 시스템(REDS)의 구현, <응용통계연구>, 제 10권, 2호, 241-251.

[3] 최용석, 현기홍 (2000). 행렬도시스템(Biplots System)의 개발, <응용통계연구>, 제 13권, 2호, 297-306.

[4] 허문열 (1995). 컴퓨터 그래픽스에 의한 이원분산분석, <응용통계연구>, 제 8권, 1호, 75-87.

[5] Jobson, J.D. (1992). *Applied Multivariate Data Analysis*, Springer-Verlag, New York, London.

[6] Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, Prentice-Hall.

[7] Rencher, A.C. (1995). *Methods of Multivariate Analysis*, John-Wiley & Sons.

[8] Shammas, N. (1996). *Mathematical Algorithms in Visual Basic for scientists and Engineers*, McGraw-Hill.

[2001년 2월 접수, 2001년 4월 채택]

Development of Statistical System for Checking Multivariate Normality and Outliers*

Yong-Seok Choi [1] Jong-Gun Kim [2] Myeong-Rae Kang [3]

ABSTRACT

Most of the techniques in multivariate analysis are based on the assumption that the data were generated from a multivariate normal distribution. So it is necessary to check a multivariate normality and outliers before analyzing multivariate data. Unfortunately, it is difficult to check multivariate normality and outliers because most statistical packages apply on only univariate case.

In this paper, firstly, we supply environment of GUI which is constructed with Visual Basic computer language. Secondly, we supplies the system to check mulitvariate normality and outliers as well as univariate case. Finally, if data set dose not satisfy the normality, we make data set nearly access normality by detecting outliers and transformation.

Keywords: Normality; Outlier; GUI; Visual Basic.

* If you want to use this system, contact http://home.pusan.ac.kr/~ yschoi.

1) Associate Professor, Department of Statistics, Pusan National University. E-mail: yschoi@hyowon.pusan.ac.kr.

2) Associate Professor, Group-department of Information Communication, Pusan College of Information Technology. E-mail: jgkim@pitc.ac.kr.

3) Graduate, Department of Statistics, Pusan National University. E-mail: banny99@hanmail.net.