한국통계학회논문집 제4권 1호
1997년 4월 pp. 41-47

다변량 자료의 분산균일성 검정 *
- 피트만 방법의 확장 -

허 명 회 1) , 양 경 숙 2)

요 약

본 연구의 목적은 p변량 관측치가 등상관구조를 갖는 경우 주변분산들의 균일성을 검정하는 통계적 절차를 개발하는 것이다. 이를 위하여 2변량의 경우에 적용되는 피트만(Pitman)의 방법을 3변량 이상의 경우로 확장하고 피셔(Fisher)의 임의화 검정을 적용하여 정규분포의 틀에 의존하지 않는 p 값을 산출한다.

1. 연구목적과 배경

n개의 p변량 관측치 $x_1, x_2, \cdots, x_n$가 등상관(等相關)구조의 다변량 정규분포 $N_p(\mu, \Sigma)$로부터 독립적으로 생성되는 경우에서 p개 변량의 분산들이 균일한가에 관심을 갖는다. 즉 $\Sigma = D_\sigma \Psi D_\sigma$, $D_\sigma = diag(\sigma_1, \cdots, \sigma_p)$, $\Psi = (1-\rho) I_p + \rho J_p$ 인 때 (I_p는 $p \times p$ 항등행렬이고 J_p는 모든 원소가 1인 $p \times p$ 행렬),

$$H_0 : \sigma_1^2 = \cdots = \sigma_p^2 (= \sigma_0^2) \quad 대 \quad H_1 : H_0 가 아님$$

의 통계적 검정법을 고안하고자 한다. 다음은 이런 검정법이 적용될 수 있는 사례들이다.

가) 논술의 채점에 있어 채점의 신뢰도를 높이기 위하여 n개의 논술답안을 p(=2)번 채점하였다. 그 결과, 첫번째 채점의 분산보다 두번째 채점의 분산이 더 작은 듯이 보였다. 이것이 채점자의 피로로 인한 변별력의 감소를 의미하는가? 아니면 단순한 임의현상일 뿐인가?

나) 통계학을 수강하는 n명의 남학생들로부터 p(=3)개 변량-아버지의 키, 어머니의 키, 본인의 키-를 조사하였다. 학생들에게 Galton이 설명하였던 방식으로 회귀현상을 보여주기 위해서였다. 그 결과, 아버지의 키, 어머니의 키, 아들의 키가 각기 다른 분산을 보여 Galton의 설명 방식을 그대로 적용하기 어려웠다. 최근 우리나라에서 영양섭취의 질적·양적 향상으로 젊은 세대의 평균키가 커진 것은 사실이다. 그런데 분산이 평균과 마찬가지로 변하는지, 아니면 분산은 평균과는 달리 변하지 않는지를 통계적으로 검정할 필요가 있다.

1) 고려대학교 정경대학 통계학과 교수. [136-701] 서울특별시 성북구 안암동 5가 1번지.
2) 고려대학교 통계학과 박사과정.
* 본연구는 1996년도 한국학술진흥재단으로부터 연구비 지원을 받아 수행되었습니다.

p개 변량이 독립 관측된 경우, 즉 Σ가 $diag(\sigma_1^2 \cdots, \sigma_p^2)$인 경우에 있어서 분산 균일성 검정은 다음과 같이 잘 알려져 있다.

· 일반화가능도비 검정 (바틀렛 검정) :

$$G = n\{p \log \bar{s}^2 - \sum_{j=1}^{p} \log s_j^2\} \sim \text{근사적 } \chi^2(p-1),$$

여기서 s_j^2는 변량 j의 표본분산이며 $(j=1, \cdots, p)$, $\bar{s}^2 = \sum s_j^2/p$.

· F-검정 (p=2인 경우) :

$$F = s_1^2 / s_2^2 \sim F(n-1, n-1).$$

p개 변량이 동시 관측된 경우, 즉 Σ가 σ_1^2, ⋯, σ_p^2를 대각원소로 하는 일반적인 공분산행렬인 경우에 대한 분산 균일성 검정의 기원은 Pitman (1939)에까지 거슬러 올라간다. 그는 p=2인 경우에 적용될 수 있는 검정법을 개발하였는데, 그 절차 및 이론은 다음과 같다 (Snedecor and Cochran (1980) 참조).

i) 원래의 변수 X_1과 X_2를 합과 차로 변환한다. 즉,

$$S = X_1 + X_2, \quad D = X_1 - X_2 .$$

ii) $Cov(S, D) = \sigma_1^2 - \sigma_2^2$ 이므로 영가설(null hypothesis)하에서는 변수 S와 D가 무상관의 관계에 놓인다. 따라서 S와 D 사이의 상관계수 r_{DS}를 구하고 이것을 0으로 볼 수 있는지에 대한 가설검정을 한다. 예컨대

$$t = \sqrt{n-2}\, r_{DS} / \sqrt{1-r_{DS}^2}, \quad d.f. = n-2$$

에 준거하여 p 값을 산출한다.

일반적으로 $p \geq 3$에 적용될 수 있는 분산 균일성에 관한 검정법으로 Harris (1985)에 의한 Wald 형의 검정법이 개발되어 있지만, Pitman의 방법은 아직 일반화되어 있지 않다.

2. 피트만 방법의 확장

p가 2인 경우에 있어 Pitman은 원래의 복합 영가설을 변수변환을 통하여 보다 단순한 영가설로 재표현할 수 있었다. 그렇다면, p가 3이상인 경우로 Pitman 방법을 확장할 수 있을까?

등상관행렬 $\Psi = (1-\rho) I_p + \rho J_p$에 고유값-고유벡터 분해를 적용하여 보자. p개 고유값이

$$\lambda_1 = 1 + (p-1)\rho, \quad \lambda_2 = \cdots = \lambda_p = 1-\rho$$

이고 해당하는 고유벡터가

$$P \equiv \begin{pmatrix} k_1 & k_2 & k_3 & \cdots & k_p \\ k_1 & -k_2 & k_3 & \cdots & k_p \\ k_1 & 0 & -2k_3 & \cdots & k_p \\ \vdots & \vdots & \vdots & & \vdots \\ k_1 & 0 & 0 & \cdots & -(p-1)k_p \end{pmatrix}, \quad \begin{matrix} k_1 = 1/\sqrt{p} \\ k_2 = 1/\sqrt{1 \cdot 2} \\ k_3 = 1/\sqrt{2 \cdot 3} \\ \vdots \\ k_p = 1/\sqrt{(p-1)\,p} \end{matrix}$$

의 각 열로 주어지므로, $\Lambda = diag(\lambda_1, \cdots, \lambda_p)$라고 할 때

$$\Psi = P \Lambda P^t$$

로 표현된다. $p \times 1$ 랜덤벡터 x를 직교행렬 P^t에 의한 선형변환

$$y = P^t x$$

에 의하여 변환하면 y의 공분산행렬은 다음과 같게 된다.

$$\Sigma_y = P^t \Sigma P = P^t D_\sigma \Psi D_\sigma P = P^t D_\sigma P \Lambda P^t D_\sigma P .$$

따라서 영가설은 $H_0^*: \Sigma_y = \sigma_0^2 \Lambda$로 재표현된다. 즉, 영가설 하에서 랜덤벡터 y의 변량간 상관계수는 모두 0이다. 우선 Λ가 기지(旣知)임을 가정하자.

$y_1, \cdots, y_n$로부터 얻게되는 2배의 로그 가능도(尤度)를 영가설(귀무가설)과 대안가설(대립가설) 하에서 비교해 보자. 영가설 하에서는

$$2 \log L_0 = -n \log | \sigma_0^2 \Lambda | - n\, tr(\Lambda^{-1} S_y) / \sigma_0^2$$

인데 (여기서 $S_y = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^t$), 이것은 $\widehat{\sigma_0}^2 = tr\{\Lambda^{-1} S_y\}/p$에서 최대가 된다. 따라서

$$\max_{\sigma_0^2} \; 2 \log L_0 = -n \log | \widehat{\sigma_0^2} \Lambda | - n p .$$

대안가설 하에서는

$$\begin{aligned} 2 \log L_1 &= -n \log | \Sigma_y | - n\, tr(\Sigma_y^{-1} S_y) \\ &= -n \log | D_\sigma^2 \Lambda | - n\, tr(D_\sigma^{-1} \Psi^{-1} D_\sigma^{-1} S_x) \end{aligned}$$

인데 (여기서 $S_x = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^t$), 이것을 최대로 하기 위하여 σ_j^2에 대하여 1, 2차 편미분하면

$$\partial \{2 \log L_1\} / \partial \sigma_j^2 = -n/\sigma_j^2 + n/\sigma_j^3 \; tr(D_\sigma^{-1} \Psi^{-1} D_j S_x),$$

$$\begin{aligned} \partial^2 \{2 \log L_1\} / (\partial \sigma_j^2 \cdot \partial \sigma_k^2) &= -n/(2\sigma_j^3 \sigma_k^3) \; tr(D_k \Psi^{-1} D_j S_x) \\ &\quad + \delta_{jk} \{ n/\sigma_j^4 - 3n/(2\sigma_j^5) \; tr(D_\sigma^{-1} \Psi^{-1} D_j S_x) \} \end{aligned}$$

이므로 (여기서 $D_j = diag(0,\cdots,1,\cdots,0)$, $j=k$일 때 $\delta_{jk}=1$, 아니면 $\delta_{jk}=0$; $j,k=1,\cdots,p$), 뉴튼-라프슨(Newton-Raphson) 방법을 적용하여 $2\log L_1$이 최대가 되는 해 $\hat{\sigma}_1^2,\cdots,\hat{\sigma}_p^2$을 수치적으로 구할 수 있다.

따라서 일반화가능도비검정(GLRT) 통계량이

$$G(\Lambda) = -n\log|D_{\hat{\sigma}}^{-2}\Lambda| - n\,tr(D_{\hat{\sigma}}^{-1}\Psi^{-1}D_{\hat{\sigma}}^{-1}S_x) + n\log|\hat{\sigma}_0^2\Lambda| + np$$

$$= -n\log(\hat{\sigma}_1^2\cdots\hat{\sigma}_p^2/\hat{\sigma}_0^{2p}) - n\,tr(D_{\hat{\sigma}}^{-1}\Psi^{-1}D_{\hat{\sigma}}^{-1}S_x) + np$$

로 주어지며 이것은 영가설하에서 점근적으로 자유도 p-1인 카이제곱분포를 따르게 될 것이다.

실제에 이 방법을 활용하기 위해서는 이제까지 기지(旣知)로 간주해둔 Λ 또는 ρ를 어떻게든 처리하지 않으면 안된다. 다음의 두 방법을 고려할 수 있을 것이다.

한 방법은 그럴듯한 범위내의 각 ρ의 값에 대하여 앞의 절차를 적용하는 것이다. 예컨대 ρ = 0.1, 0.2, 0.3, 0.4, 0.5의 각각에 대하여 GLRT 통계량 G를 계산하고 그 변화가 얼마나 민감한가를 보는 것이다. 또 하나의 방법은 ρ를 점근일치성을 갖는 직관적인 추정치인

$$\bar{r} = \sum_{i=1}^{p}\sum_{j=i+1}^{p} r_{ij}/\{p(p-1)/2\}$$

로 대치시키는 것이다. 여기서 r_{ij}는 x내 i번째 변량과 j번째 변량간의 상관계수이다.

3. 피셔의 임의화 검정

이제까지는 다변량 정규분포가 가정되었다. 그러나 피셔(R.A. Fisher)의 임의화 검정(Fisher randomization test; FRT)은 그런 가정 없이 가능하다는 장점이 있다. 이에 관한 역사적 기원에 관하여는 Fisher (1966)를, 일반적 이론과 응용에 관하여는 Good (1994)을 참조할 수 있다. 다변량관측의 분산균일성을 위한 검정에서 FRT에 의한 p 값의 계산방법은 다음과 같이 유도될 수 있다.

x의 변량들이나 y의 변량들은 영가설 하에서도 각기 다른 분산을 가지므로 호환가능하지 않다. 그렇지만

$$z = \Lambda^{-1/2}y = \Lambda^{-1/2}P^t x$$

로 변환하면 세 벡터는 $\sigma_0^2 I_p$를 공분산행렬로 갖게 되므로 중심화된 z의 변량들은 서로 호환가능하게 된다.

z의 1개 관측당 $p!$개의 호환가능한 경우가 생기며 관측 수가 n인 경우 $p!^n$ 개 총가능자료의 확률적 균등성을 바탕으로 FRT를 구성할 수 있다. 단, 총가능 수가 매우 크므로 모든 가능자료로부터 검정통계량을 각각 계산하는 것은 무리이다. 그러나 p 값의 산출을 목적으로 하는 경우 영가설 하에서 검정통계량의 확률분포를 유도하면 되므로 모든 가능자료의 일부를 몬테칼로(Monte Carlo) 추출하는 것만으로 족하다.

4. 수치예

여기서 수치예로 들려는 자료는 연구자들이 수집한 K대 통계학과 남학생 64명의 부·모와 본인의 키 자료 (p=3)이다. 이 자료에서 아버지의 키(=x_1), 어머니의 키(=x_2), 학생의 키(=x_3)의 표본분산과 변량간 상관계수는 다음과 같다.

표본분산 s_1^2 = 28.2, s_2^2 = 17.1, s_3^2 = 34.0,
상관계수 r_{12} = 0.17, r_{13} = 0.27, r_{23} = 0.24. (평균 0.23)

여기서 변량들이 등상관구조를 갖는 것으로 전제하고 주변분산들이 균일한가를 검정하기로 하자. ρ의 5개 값인 0.15, 0.20, 0.23, 0.25, 0.30의 각각에 대하여 GLRT 통계량 G를 계산하고 정규분포하에서의 근사적 p 값과 FRT에 의한 p 값을 계산해 보았다. FRT에서 반복수는 10,000번이었고 따라서 p 값에 대한 95% 신뢰수준의 오차한계는 ±0.006 이내이다. 그 결과는 <표 1>과 같다. FRT의 p 값이 정규분포하에서의 근사적 p 값보다 대체로 큰 것으로 나왔는데 이는 자료에 <그림 1>에서 볼 수 있는 바와 같이 1개의 뚜렷한 특이점이 있기 때문인 것으로 생각된다. 만약 그 관측치를 분석자료에서 제외하는 경우에는

표본분산 s_1^2 = 27.7, s_2^2 = 12.9, s_3^2 = 34.1,
상관계수 r_{12} = 0.31, r_{13} = 0.26, r_{23} = 0.34 (평균 0.30)

이 된다. <표 2>는 ρ의 3개 값 0.25, 0.30, 0.35 각각에 대하여 GLRT 통계량 G를 계산하고 정규분포하에서의 근사적 p 값과 FRT에 의한 p 값을 다시 계산한 결과이다. 앞에서와 마찬가지로 FRT에서 반복수는 10,000번이었다. GLRT 통계량을 대표본 정규근사론에서 보느냐, 또는 피셔의 임의화 틀에서 보느냐에 따라 p 값이 달라지기는 해도 그럴듯한 등상관계수 ρ의 범위내에서 자체적으로는 거의 변하지 않음을 볼 수 있다.

5. 마치며

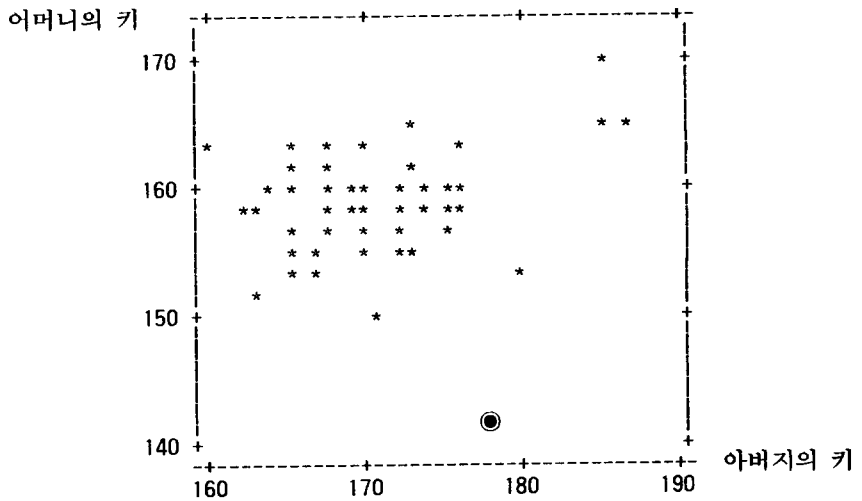
본 연구교신은 분산균일성에 관한 피트만의 검정 방법을 확장하는 것이었고 내용상으로는 변수변환을 통해 일반화가능도비검정(GLRT) 통계량 G를 유도하고 뉴튼-라프슨 알고리즘을 통해 계산해낸 것이다. 이것은 선행연구인 Harris (1985)가 개발한 Wald형의 검정통계량 W와 소표본에서의 검정의 크기 및 검정력(power) 등 통계적 측면에서 비교될 수 있을 것이다. 이에 관하여는 추후의 연구로 넘기기로 하겠다. G와 W의 비교에 관한 일반적인 연구 결과에 의하여(예컨대, Buse (1982)), 이들이 대표본에서는 일치하지만 소표본에서는 W의 검정력이 G에 비하여 상대적으로 작은 경향이 있을 것으로 예상할 수 있겠다.

<표 1> 등상관구조하에서의 분산 균일성 검정 결과

등상관계수 ρ	max 2 log L_1	max 2 log L_0	GLRT 통계량 G	근사 p 값 *	FRT p 값**
0.15	-804.912	-812.626	7.714	0.0211	0.1379
0.20	-804.091	-811.826	7.735	0.0209	0.0307
0.23	-803.980	-811.728	7.748	0.0208	0.0963
0.25	-804.058	-811.814	7.756	0.0207	0.0626
0.30	-804.770	-812.546	7.776	0.0205	0.0667

* 자유도 2(=p-1)인 카이제곱 분포하에서의 유의확률 ** 피셔의 임의화 검정에 의한 유의확률

<그림 1> 아버지와 어머니의 키



<표 2> 특이점 제거 후의 분산 균일성 검정 결과

등상관계수 ρ	max 2 log L_1	max 2 log L_0	GLRT 통계량 G	근사 p 값 *	FRT p 값**
0.25	-766.273	-783.286	17.013	0.0002	0.0028
0.30	-765.836	-783.202	17.366	0.0002	0.0026
0.35	-766.155	-783.863	17.708	0.0001	0.0024

* 자유도 2(=p-1)인 카이제곱 분포하에서의 유의확률 ** 피셔의 임의화 검정에 의한 유의확률

참고문헌

[1] Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier test: An expository note, *American Statistician* 36, 153-157.

[2] Fisher, R.A. (1966, 1st edition 1935). *The Design of Experiments*, 8th Edition, Oxford University Press, 11-26.

[3] Good, P. (1994). *Permutation Tests,* Springer Verlag, New York.

[4] Harris, P. (1985). Testing for variance homogeneity of correlated variables, *Biometrika* 72, 103-107.

[5] Pitman, E.J.G. (1939). A note on normal correlation, *Biometrika* 31, 9-12.

[6] Snedecor, G.W. and Cochran, W.G. (1980). *Statistical Methods*, 7th Edition, Iowa State University Press, 190-191.