The Korean Communications in Statistics
Vol. 13 No. 2, 2006, pp. 429-440

Distribution of a Sum of Weighted Noncentral Chi-Square Variables[1)]

Sunyeong Heo[2)] and Duk-Joon Chang[3)]

Abstract

In statistical computing, it is often for researchers to need the distribution of a weighted sum of noncentral chi-square variables. In this case, it is very limited to know its exact distribution. There are many works to contribute to this topic, e.g. Imhof (1961) and Solomon-Stephens (1977). Imhof's method gives good approximation to the true distribution, but it is not easy to apply even though we consider the development of computer technology. Solomon-Stephens's three moment chi-square approximation is relatively easy and accurate to apply. However, they skipped many details, and their simulation is limited to a weighed sum of central chi-square random variables. This paper gives details on Solomon-Stephens's method. We also extend their simulation to the weighted sum of non-central chi-square distribution. We evaluated approximated powers for homogeneous test and compared them with the true powers. Solomon-Stephens's method shows very good approximation for the case.

Keywords : Newton-Raphson iteration; Wald test; Homogeneous test.

1. 서론

$L\times 1$ 확률벡터 $X\sim N_L(\mu,\Sigma)$이고 Σ가 정칙행렬일 때, 일반적으로 $Q=X'AX$로 표현되는 2차형식(quadratic forms)은 다음과 같이 자유도가 1인 비중심카이제곱확률변수들의 가중합으로 표현할 수 있다:

$$Q=\sum_{k=1}^{L} c_k(Z_k+a_k)^2. \tag{1.1}$$

이 때, Z_k는 표준정규분포를 따르는 확률변수로 상호 독립이고 동일한 분포를 갖으며, c_k와 a_k는 각각 음이 아닌 상수이다. 식(1.1)의 k번째 항 $U_k=Z_k+a_k$는 $U_k\sim N(a_k,1)$이므로 U_k^2은 자유도가 1이고 비중심모수가 a_k^2인 비중심카이제곱분포를 따른다.

1950년대 이후 연구자들은 식(1.1)로 표현되는 이차형식의 분포에 관심을 가지고 활

1) This research is financially supported by Changwon National University in 2005.
2) Assistant Professor, Dept. of Statistics, Changwon National University, Changwon, Korea.
3) Professor, Dept. of Statistics, Changwon National University, Changwon, Korea.
Correspondence : djchang@sarim.changwon.ac.kr

발한 연구를 진행했으나, 일반적으로 이러한 형태의 이차형식의 정확한 분포를 계산할 수 있는 경우는 극히 제한되어 있다. 초기 연구들은 주로 모든 k에 대해 $a_k=0$인 중심카이제곱분포의 선형결합에 관한 것으로 Gurland (1953, 1954), Box (1954) 등이 그 경우이다. 그들은 $a_k=0$일 때 Q의 구체적인 분포형태와 여러 가지 근사방법들을 제시하고 있다.

Tiku (1985)는 모든 k에 대해 $c_k=1$일 때 Q의 확률밀도함수, 분포함수, 적률모함수의 구체적인 형태와 분포함수에 대한 세 가지 근사방법을 비교·소개하고 있다. 이 경우의 분포함수 형태와 근사방법에 대해서 Kerridge (1965), Fraser et. al. (1998), Posten (1989)을 참고할 수 있다.

Imhof (1961)는 $a_k\neq 0$이고 $c_k\neq 1$인 보다 일반적인 형태의 Q의 분포함수와 근사식을 시뮬레이션 결과와 함께 보여주고 있다. 또, 비중심카이제곱 분포에 대한 Pearson (1959)의 근사식을 Q의 근사분포에 적용한 근사식과 시뮬레이션 결과도 함께 보여준다. 이 경우에 대한 또 다른 근사방법으로 Jensen과 Solomon (1972)은 Q의 분포를 정규근사시키는 방법과 시뮬레이션 결과를, Solomon과 Stephens (1977)는 카이제곱분포에 근사시키는 방법과 시뮬레이션 결과를 보여준다.

Solomon & Stephens (1977)가 제시한 3차적률카이제곱근사법은 컴퓨터의 발달과 더불어 실제 분석에 용이하게 적용할 수 있는 방법으로, 시뮬레이션을 통해 모든 k에 대해 $a_k=0$인 경우 3차적률카이제곱근사법이 참 분포에 매우 근사함을 보여주고 있다.

본 연구는 Solomon-Stephens가 제시한 3차적률카이제곱근사법의 구체적인 수식을 유도하고(2절), 동질성검정의 Wald 검정통계량의 검정력 계산을 통해 $a_k\neq 0$인 경우 Solomon-Stephens의 3차적률카이제곱근사법이 얼마나 참값에 근사한 결과를 제공하는지를 시뮬레이션 결과를 통해 확인하였다(3절).

2. Solomon-Stephens의 3차적률카이제곱근사법

2.1 비중심카이제곱 확률변수들의 선형결합의 분포

확률변수 U가 자유도 p, 비중심모수가 λ인 비중심카이제곱분포를 따를 때, 즉, $U\sim\chi^2(p,\lambda)$일 때, $W=cU$ ($c>0$인 상수)의 적률모함수, $M_W(t)$,는

$$M_W(t)=(1-2ct)^{-p/2}\exp\left(\frac{c\lambda t}{1-2ct}\right)\quad,\quad t<\frac{1}{2c}$$

이고, 적률모함수의 정의에 의해 식(1.1)의 Q의 적률모함수, $M_Q(t)$,는

$$M_Q(t)=\prod_{k=1}^{K}(1-2c_kt)^{-1/2}exp\left(\frac{c_k\lambda_k t}{1-2c_kt}\right),\quad t<\frac{1}{2c_{\max}} \tag{2.1}$$

가 된다. 이 때, $\lambda_k=a_k^2$이고 $c_{\max}$는 $(c_1,\cdots,c_L)$중에서 가장 큰 수이다.

$B(t)=\sum_{k=1}^{L}c_k/(1-2c_kt)+\lambda_kc_k/(1-2c_kt)^2$라 할 때, 식(2.1) $M_Q(t)$의 1, 2, 3차 도

함수는 각각

$$M_Q^{(1)}(t) = B(t) \cdot M_Q(t),$$
$$M_Q^{(2)}(t) = B^{(1)}(t) + (B(t))^2 M_Q(t)$$
$$M_Q^{(3)}(t) = B^{(2)}(t) + 3 \cdot B^{(1)}(t) \cdot B(t) + (B(t))^3 M_Q(t)$$

로 표현된다. 이 때, $B^{(n)}(t)$는 $B(t)$의 n차 도함수를 나타낸다. 또, 적률과 적률모함수의 관계에 의해

$$\begin{aligned} \mu &= E(Q) = M_Q^{(1)}(0) = B(0), \\ \mu_2' &= E(Q^2) = M_Q^{(2)}(0) = B^{(1)}(0) + (B(0))^2, \\ \mu_3' &= E(Q^3) = M_Q^{(3)}(0) = B^{(2)}(0) + 3 \cdot B^{(1)}(0) \cdot B(0) + (B(0))^3 \end{aligned} \tag{2.2}$$

가 되고, 여기서

$$\begin{aligned} B(0) &= \sum_{k=1}^{L} c_k (1 + a_k^2), \\ B^{(1)}(0) &= 2 \sum_{k=1}^{L} c_k^2 (1 + 2a_k^2), \\ B^{(2)}(0) &= 8 \sum_{k=1}^{L} c_k^3 (1 + 3a_k^2) \end{aligned} \tag{2.3}$$

이다. Solomon-Stephens의 정의에 따라 $R_2 = \mu_2'/\mu^2$ 와 $R_3 = \mu_3'/\mu^3$라 놓으면, 식 (1.1) Q의 R_2와 R_3은 식(2.2) - 식(2.3)으로부터 다음과 같이 계산될 수 있다:

$$\begin{aligned} R_2 &= \frac{B^{(1)}(0) + (B(0))^2}{(B(0))^2} \\ R_3 &= \frac{B^{(2)}(0) + 3B^{(1)}(0) \cdot B(0) + (B(0))^3}{(B(0))^3} \end{aligned} \tag{2.4}$$

2.2 Solomon-Stephens에 의한 근사식

확률변수 X가 자유도 p인 카이제곱분포를 따를 때, 즉 $X \sim \chi^2(p)$일 때, 임의의 양의 상수 A와 r에 대해 $Q_s = AX^r$라 하자. 확률변수 X의 확률밀도함수와 변수변환에 의해, $Y = X^r$의 확률밀도함수는

$$f(y) = \frac{1}{\Gamma(v) \cdot 2^v \cdot r} y^{(v/r)-1} \exp\left(-\frac{y^{1/r}}{2}\right), \quad y > 0 \tag{2.5}$$

이고, 이 때 $v = p/2$이다. 식(2.5)와 기대값의 정의로부터 확률변수 Y의 m차 적률은

$$E(Y^m) = 2^{mr} \cdot \Gamma(mr + v)/\Gamma(v), \quad m = 1, 2, 3, \cdots$$

이 되고, 그 결과 Q_s의 m차 적률은

$$E(Q_s^m) = A^m \cdot E(Y^m) = A^m \cdot 2^{mr} \cdot \Gamma(mr + v)/\Gamma(v)$$

이므로, Q_s의 기대값은

$$\mu = E(Q_s) = A \cdot E(Y) = A \cdot 2^r \cdot \Gamma(r+v)/\Gamma(v) \tag{2.6}$$

이고, Q_s의 2차와 3차 적률들은 각각

$$\mu_2' = E(Q_s^2) = A^2 \cdot E(Y^2) = A^2 \cdot 4^r \cdot \Gamma(2r+v)/\Gamma(v),$$
$$\mu_3' = E(Q_s^3) = A^3 \cdot E(Y^3) = A^3 \cdot 8^r \cdot \Gamma(3r+v)/\Gamma(v)$$

이다.

Solomon-Stephens가 제안한 3차적률카이제곱근사법에 따라 식(1.1) Q의 분포가 $Q_s = AX^r$의 분포와 근사적으로 같다고 가정하면, 식(2.4)의 (R_2, R_3)은 Q_s의 적률들로부터 유도된 (R_2, R_3)에 근사한 값을 갖게 될 것이다. 따라서 그것들을 같다고 두면

$$R_2 = \frac{\Gamma(v) \cdot \Gamma(2r+v)}{\Gamma(r+v)^2}$$
$$R_3 = \frac{\Gamma(v)^2 \cdot \Gamma(3r+v)}{\Gamma(r+v)^3} \tag{2.7}$$

이 된다. 이 때, 식(2.7)의 왼쪽의 (R_2, R_3)은 식(2.4)에서 계산된 값이다. 식(2.7)은 감마함수를 포함한 (v,r)의 비선형방정식이다. 식(2.7)을 (v,r)에 대해 해를 구한 후, 그 해와 식(2.2)의 $\mu = B(0)$을 식(2.6)에 대입하여 A의 값을 얻을 수 있다. 이렇게 얻은 (v,r,A)로부터 식(1.1) Q의 분포는 $P(Q < t) \approx P(AX^r < t)$이 된다.

3. Solomon-Stephens의 3차적률카이제곱근사법을 이용한 동질성검정의 Wald 통계량의 검정력

Solomon과 Stephens (1977)는 Q_s의 분포가 Q의 분포에 근사함을 경험적으로 보여주기 위해 시뮬레이션 결과를 이용하고 있다. 시뮬레이션 과정에서는 모든 k에 대해 $a_k = 0$을 가정하였다. 즉, 그들은 중심카이제곱확률변수들의 선형결합의 근사분포에 대한 결과만을 제시한다. 우리는 $a_k \neq 0$인 경우, 3차적률카이제곱근사법에 의한 확률이 참값에 근사한 정도를 확인하기 위해서 동질성검정을 위한 Wald 검정통계량의 검정력을 계산하였다.

3.1 동질성검정의 Wald 검정통계량의 검정력

서로 독립인 두 모집단으로부터 크기가 각각 $n_i (i = 1,2)$인 표본을 추출하고, 상호배반인 K개의 범주에 속할 확률이 동일한가를 검정할 때, 각 범주에 속할 표본비율들의 분산을 알 수 있다면 Wald 통계량을 통한 검정을 한다. 즉, i번째 모집단의 모비율을 p_i, 표본비율을 $\hat{p}_i$, 그리고 $\hat{p}_i$의 분산을 V_i라 하자. 이 때, p_i와 $\hat{p}_i$은 처음 $K-1$개 범주에 속할 비율을 나타내는 $(K-1)\times 1$ 벡터이다. 이 경우, 두 모집단의 동질성, $H_0 : p_1 = p_2 (= p)$을 검정하기 위한 Wald 검정통계량은

$$X_W^2 = (\hat{p_1} - \hat{p_2})'(V_1 + V_2)^{-1}(\hat{p_1} - \hat{p_2})$$

이다. 실제로 많은 경우에는 V_i를 알지 못하므로 V_i의 일치추정값을 사용한다.

표본비율 $\hat{p_i}$의 근사분포가 $N(p_i, V_i)$라면, $(\hat{p_1} - \hat{p_2})$의 분포도 근사적으로 정규분포 $N(p_1 - p_2, V_1 + V_2)$가 된다. 또, X_W^2의 분포는 근사적으로 자유도가 $K-1$이고 비중심모수가 $\lambda = (p_1 - p_2)'(V_1 + V_2)^{-1}(p_1 - p_2)$인 카이제곱분포를 따른다(Graybill, 1976, p.127).

여기서, 자유도가 $K-1$이고 비중심모수가 $\lambda = (p_1 - p_2)'(V_1 + V_2)^{-1}(p_1 - p_2)$인 카이제곱분포를 갖는 확률변수를 Q_W라 하자. 유의수준이 α일 때, $H_1 : p_1 \neq p_2$하에서 Q_W의 검정력은

$$1 - \beta_W = \Pr\{Q_W > \chi^2_{K-1,\alpha} | D = p_1 - p_2\} \tag{3.1}$$

로 표현되고, $\chi^2_{K-1,\alpha}$는 자유도가 $K-1$인 카이제곱분포의 상위 $100\alpha\%$ 백분위수이다. 식(3.1)의 $1-\beta_W$는 X_W^2의 근사검정력이 된다. 여기서 Q_W의 분포는 비중심모수 λ의 값에 따라 달라지고, λ는 두 모비율의 차이 $D = p_1 - p_2$와 표본비율들의 분산 (V_1, V_2)에 의해서 결정되므로, 식(3.1)에서 검정력으로 표현되는 확률 $1-\beta_W$를 계산하기 위해서는 이들의 값들을 알아야 한다.

비중심모수가 0이 아닌 카이제곱확률변수들의 선형결합에 대한 Solomon-Stephens의 3차적률카이제곱근사법을 적용하기위해 Q_W를 비중심카이제곱 확률변수들의 선형결합으로 표현할 필요가 있다.

Corollary 1. $L \times 1$ 확률벡터 X의 분포가 $N_L(\mu, \Sigma)$이고, Σ는 정칙행렬이라 하자. 이 때 2차형식 $Q = X'\Sigma^{-1}X$는 다음과 같이 표현할 수 있다:

$$Q = \sum_{k=1}^{L} (Z + \lambda_k)^2.$$

이 때 $Z \sim N(0,1)$이고 λ_k은 $\lambda = \Sigma^{-1/2}\mu$의 k번째 원소이다.

증명. $Q = X'\Sigma^{-1}X = (\Sigma^{-1/2}X)'(\Sigma^{-1/2}X) = \sum_{k=1}^{L} Z_k^2$, 여기서 Z_k은 $\Sigma^{-1/2}X$의 k번째 원소이다. $\Sigma^{-1/2}X$의 분포는 $N_L(\Sigma^{-1/2}\mu, I)$이므로, Z_k은 상호독립이고 $N(\lambda_k, 1)$인 분포를 따른다. $Z \sim N(0,1)$라 할 때, $Z_k = Z + \lambda_k$로 나타낼 수 있다. 또, Z_k^2은 자유도가 1이고 비중심위치모수가 λ_k^2인 비중심카이제곱분포를 따른다. 즉, $Z_k^2 \sim \chi^2(1, \lambda_k^2)$. □

Corollary 1에 의해서 Q_W를 다음과 같이 표현할 수 있다:

$$Q_W = \sum_{k=1}^{K-1} Z_k^2 = \sum_{k=1}^{K-1} (Z + \mu_k)^2. \tag{3.2}$$

이 때, $Z \sim N(0,1)$이고 $\mu_k = (V_1 + V_2)^{-1/2}(p_1 - p_2)$의 k번째 원소이다. 식(3.2)의 Q_W는 식(1.1)의 특별한 경우로, 모든 k에 대해 $c_k = 1$이고, $a_k = \mu_k$인 경우이다.

식(3.1)에 대한 Solomon-Stephens의 3차적률카이제곱근사법에 의한 근사확률을 얻기 위해, 식(2.3)에 $((c_k, a_k) = (1, \mu_k)$를 대입하여 식(2.2)와 식(2.4)로부터 μ와 (R_2, R_3)를 계산하고, 그 값을 식(2.7)에 대입하여 (v, r)의 해를 구한 후, 다시 식(2.6)으로부터 A를 계산한다. 그 결과로 얻은 (r, v, A)를 대입하여

$$1 - \beta_W \approx \Pr\{AX^r > \chi^2_{K-1,\alpha} | D = p_1 - p_2\} \tag{3.3}$$

를 얻는다. 여기서 $X \sim \chi^2(p)$, $p = 2v$이다.

3.2 Solomon-Stephens의 3차적률카이제곱근사를 이용한 검정력계산

앞 절에서 언급한 바와 같이 식(3.1)과 식(3.3)에서 검정력으로 표현된 확률들을 계산하려면 두 표본비율들의 분산 (V_1, V_2)와 두 모비율간의 차이 $D = p_1 - p_2$를 알 필요가 있다. 우리는 두 표본비율들의 분산을 얻기 위해, 미국국민건강면접조사의 자료를 이용하여 분산을 추정한 후, 그 값을 참값으로 사용하였다. 두 모비율간의 차이를 얻기 위해, 필요한 조건을 부여한 후 그 조건을 만족하는 값들을 균등분포를 이용하여 임의추출하였다.

3.2.1 미국국민건강면접조사

우리는 분산추정값을 얻기 위해 미국의 2003국민건강면접조사(U.S. 2003 National Health Interview Survey: U.S. NHIS)의 인터넷 자료를 이용하였다. 미국의 국민건강면접조사는 미국의 어린이들을 포함한 일반시민들의 건강상태를 파악하기 위해 매해 실시되는 조사로 다단계층화확률표본추출법에 의해 표본을 추출한다. 인터넷을 통해서 제공되는 자료는 개인을 보호하기 위해 응답자들의 소재지에 대한 구체적인 정보를 제공하지 않고, 단지 미국 전역을 네 개의 지역으로 구분하고 있다: 북동부, 중서부, 남부, 그리고 서부. 우리는 북동부와 중서부를 각각 독립된 모집단으로 간주 하였다. 즉, 북동부를 첫 번째 모집단, 중서부를 두 번째 모집단으로 보고 추정된 $\hat{V}_1 + \hat{V}_2$을 참값인 $V_1 + V_2$로 간주하였다. 분석에 사용된 각 범주형 변수의 분산은 Stata 프로그램으로부터 선형화방법(linearization method)을 사용하여 추정하였다.

3.2.2 검정력의 계산

K개의 범주를 갖는 임의의 범주형변수에 대해, 서로 독립인 두 모집단의 분포가 동일하다면, 즉 $H_0 : p_1 = p_2 (= p)$가 참이라면, $D = p_1 - p_2 = 0$이고 두 벡터 p_1과 p_2의 거리는 0이 된다. 이 때, 범주수가 K인 경우이므로, p_1과 p_2는 각각 $(K-1) \times 1$ 벡터이다.

동질성검정에서 검정력은 $D=p_1-p_2\neq 0$인 경우, 귀무가설을 기각할 확률로 정의된다. 범주수가 K인 경우, $D=p_1-p_2\neq 0$인 점들은 $(K-1)$차원 공간에서 무수히 많은 점들로 표현될 것이다. 그러나 두 모집단의 분포가 완전히 일치하지 않지만 매우 유사하다면, 두모비율 간의 거리는 0에 가까운 값이 될 것이며, 두 모집단의 분포가 전혀 다르다면 두 모비율간의 거리는 큰 값을 가질 것이다.

두 모비율 p_1과 p_2의 거리는

$$|p_1-p_2|=\sqrt{(p_{11}-p_{21})^2+(p_{12}-p_{22})^2+(p_{1,K-1}-p_{2,K-1})^2},$$

로 정의된다. $H_0: p_1=p_2(=p)$가 참일 때 $|p_1-p_2|=0$이고, $H_1: p_1\neq p_2$이 참일 때는 $0<|p_1-p_2|\leqq\sqrt{K-1}$이다. 여기서 p_{ik}는 i번째 모집단에서 k번째 범주에 속할 비율이다.

검정력의 성질에 의해서, 두 모수가 매우 유사하다면 비록 귀무가설이 거짓이라도 귀무가설을 기각할 확률은 작을 것이며, 두 모수가 크게 상이하다면 귀무가설을 기각할 확률은 크게 된다. 이러한 성질에 의해서, 우리는 균등분포 $U(0,1)$을 통해서 $D=p_1-p_2\neq 0$을 만족하는 D를 시뮬레이션하는 과정에서 두 모비율간의 거리를 일정하게 한 후, 동일한 거리를 갖는 D를 100회 반복해서 추출하였다. 즉, $0<l\leqq\sqrt{K-1}$인 서로 다른 l값에 대해 $|p_1-p_2|=l$을 일정하게 한 후, $0<|p_1-p_2|\leqq 1$을 만족하는 벡터 D를 100회 반복추출하였다.

벡터 D를 100회 반복추출하는 과정에서 먼저 우리는 미국의 2003국민건강면접조사에서 선택한 변수들에 대해, 북동부지역에 대해 표본추출에 기초한 표본비율을 추정한 후, 추정된 비율을 북동부지역의 모비율로 간주하였다. (여기서 추정된 처음 $(K-1)$개 범주들의 비율들로 이루어진 벡터를 a_1이라 하자). 그 다음, 균등분포 $U(0,1)$로부터 K개 값들을 랜덤추출하여 그 결과를 중서부지역의 모비율로 놓았다. (이 경우에도 처음 $(K-1)$개의 값만을 포함하는 벡터를 a_2라 놓자.) 임의의 $l\in l: 0<l\leqq\sqrt{K-1}$에 대해 $|p_1-p_2|=l$ 이 되는 $D=p_1-p_2$을 얻기 위해, 주어진 l에 대해

$$a_{1k}\geqq a_{2k}\ \text{이면}\ d_k=p_{1k}-p_{2k}=\sqrt{(l/|a_1-a_2|)^2(a_{1k}-a_{2k})^2},$$

$$a_{1k}<a_{2k}\ \text{이면}\ d_k=p_{1k}-p_{2k}=-\sqrt{(l/|a_1-a_2|)^2(a_{1k}-a_{2k})^2},$$

로 정의했다. 여기서 a_{1k}, a_{2k}, d_k는 각각 a_1, a_2, D의 k번째 원소로, $k=1,\cdots,K-1$이다.

이러한 과정을 100회 반복해서 100개의 벡터 D를 얻는 과정에서 서로 다른 l에 대해 D의 값들은 달라지나, 처음의 a_1과 a_2는 모든 반복에 대해 동일하다.

이러한 과정을 통해 추출된 100개의 반복된 D에 대해 식(3.1)에 의한 정확한 확률과 식(3.3)에 의한 근사확률을 구했다. 이 과정에서 식(2.7)의 해는 Newton-Raphson 방법을 사용하였다.

Newton-Raphson에 의해서 식(2.7)의 (r,v)의 해를 구하는 과정에서, l이 큰 값을 가지는 경우 R_2와 R_3가 매우 작은 값을 가지게 되면서 식(2.7)을 만족하는 해를 구할

수 없었다(예를 들어, $R_2 < 1.006$, $R_3 < 1.015$인 경우). 그러나 이 경우 그보다 작은 l 값에서 이미 검정력이 1이 되었으므로 검정력의 성질에 의해, 식(2.7)의 해를 구할 수 없고 그 결과 근사 확률을 계산할 수 없는 벡터 D에 대해서는 식(3.3)의 근사확률에 모두 1의 값을 주었다. 예를 들어, <표 3.1>에서 norm=0.030 (즉, $l = 0.030$)인 경우 100개의 벡터 D중에서 1개의 벡터에 대해 수렴하는 (r, v)를 얻지 못했다. 그러나 해를 얻지 못한 벡터 a_2에 대해 $l \leq 0.020$인 경우 검정력이 이미 1이었으므로 $l = 0.030$에 대해서도 1의 값을 주었다. (모든 l에 대해 D는 다르나 a_1과 a_2는 동일함에 주의). 이와 동일한 규칙을 <표 3.2> ~ <표 3.5>에도 적용하였다. 표에 그러한 반복을 갖는 "norm"에 "*" 또는 "**"를 표시한 후, 주에 해당하는 반복의 수를 제시하였다.

<표 3.1> ~ <표 3.5>는 범주수가 각각 4, 6, 8일 때, 정확한 검정력과 근사검정력을 계산한 결과들을 보여준다. 각 표에서 첫 번째 열의 "norm"은 두 벡터간의 거리, $|p_1 - p_2|$,를 나타낸다. 두 번째 열의 "정확한 확률"은 식(3.1)에 의해 100회 반복 계산한 검정력들의 평균을, 세 번째 열의 "근사 확률"은 식(3.3)에 의해 100회 반복 계산한 검정력들의 평균을 보여준다. "정확한 확률-근사 확률"은 각 반복으로부터 두 검정력의 차이를 계산하고, 그 100개 값에 대해 요약한 통계량이다: 네 번째 열의 "평균"은 100개의 검정력들 간의 차이들의 평균; 다섯 번째 열은 그것들의 표준편차; 여섯 번째 열은 최소값; 마지막 열은 최대값을 각각 나타낸다.

<표 3.1>과 <표 3.2>는 범주수는 $K = 4$로 동일하나, 서로 다른 두 변수들에 대해 계산된 결과이다. 참고로, 각 표에 추정된 분산-공분산행렬 $\hat{V}_1 + \hat{V}_2$을 첨부하였다. <표 3.1>은 변수명 LNG_INTV에 대한 결과로 면접조사시 사용한 언어에 관한 문항, <표 3.2>는 변수명 FSPEDCT에 대한 결과로 각 가구에서 특수교육을 받는 18세 미만 어린이들의 숫자를 묻는 문항이다. <표 3.3>과 <표 3.4>는 $K = 6$인 경우로, <표 3.3>은 변수명이 FM_SIZE로 가구당 인원수를 묻는 문항, <표 3.4>는 변수명 FWRKLWCT에 대한 결과로 조사가 이루어지기 전 한주동안 완전고용상태에 있었던 가구원 수를 묻는 문항이다. <표 3.5>는 변수명 FHICOST에 대한 결과로 과거 1년간 지출한 의료비에 관한 문항이다.

<표 3.1> ~ <표 3.5>에 제시된 모든 결과는 Solomon-Stephens의 3차적률카이제곱근사법에 의한 근사검정력이 참값과 매우 근소한 차이를 가짐을 보여준다.

4. 결론

통계계산에서 비중심카이제곱확률변수들의 선형결합으로 표현되는 확률변수의 분포를 필요로 하는 경우가 종종 있다. 예로 Rao-Scott (1981)은 적합도검정이나 독립성검정에서 표본비율의 공분산을 모르는 경우 Pearson 검정통계량에 대한 조정된 통계량을 제시하였고, 조정된 통계량은 중심 또는 비중심 카이제곱확률변수들 가중합으로 표현된다.

Solomon-Stephens (1977)는 비중심카이제곱변수들의 가중합의 분포를 단일 카이제곱확률변수의 특정 함수형태로 표현되는 분포에 근사시키는 방법을 제시하고 있다.

<표 3.1> $K=4$일 때, 정확한 검정력과 근사 검정력 및 분산추정치 (변수명: LNG_INTV)

$$\hat{V}_1+\hat{V}_2=\begin{bmatrix} 8.600E-06 & & \\ -3.636E-06 & 2.706E-06 & \\ -3.081E-06 & 1.019E-06 & 2.137E-06 \end{bmatrix}$$

norm	정확한 확률	근사 확률	정확한 확률 - 근사 확률			
			평균	표준편차	최소	최대
0	0.05	0.05	0.05	0	0.05	0.05
0.001	0.058160	0.058152	8.727E-06	7.481E-06	4.137E-06	4.924E-05
0.002	0.084766	0.084722	4.405E-05	2.593E-05	2.458E-05	2.347E-04
0.004	0.205156	0.205031	1.248E-04	3.550E-04	-1.804E-05	1.997E-03
0.006	0.405315	0.404096	1.997E-03	9.305E-04	-9.030E-04	3.111E-03
0.008	0.634044	0.631704	2.340E-03	9.188E-04	-1.003E-03	3.123E-03
0.010	0.819217	0.817720	1.497E-03	1.553E-03	-1.278E-03	-1.278E-03
0.015	0.989495	0.990273	-7.784E-04	5.145E-04	-1.278E-03	0
0.020	0.999899	0.999941	-4.195E-05	4.503E-05	-1.375E-04	0
0.030*	1	1	-2.164E-11	3.826E-11	-1.650E-10	0

* 1개 반복에서 근사값을 구하지 못함.

<표 3.2> $K=4$일 때, 정확한 검정력과 근사 검정력과 분산추정치 (변수명: FSPEDCT)

$$\hat{V}_1+\hat{V}_2=\begin{bmatrix} 0.00009597 & & \\ -0.00007682 & 0.00007528 & \\ -0.00001592 & 6.432E-07 & 0.00001527 \end{bmatrix}$$

norm	정확한 확률	근사 확률	정확한 확률 - 근사 확률			
			평균	표준편차	최소	최대
0	0.05	0.05	0.05	0	0.05	0.05
0.002	0.056251	0.056251	6.300E-06	9.177E-06	3.752E-07	4.732E-05
0.004	0.076809	0.076778	3.077E-05	7.550E-05	-1.602E-05	7.613E-04
0.006	0.113940	0.113851	8.933E-05	3.559E-04	-1.804E-05	3.105E-03
0.008	0.166329	0.166141	1.883E-04	5.258E-04	-1.499E-04	3.070E-03
0.010	0.230557	0.230162	3.945E-04	7.828E-04	-1.102E-03	3.102E-03
0.015	0.415435	0.414541	8.939E-04	1.095E-03	-1.273E-03	3.123E-03
0.020	0.589407	0.588308	1.099E-03	1.474E-03	-1.268E-03	3.123E-03
0.025	0.729010	0.727883	1.127E-03	1.354E-03	-1.276E-03	3.122E-03
0.030	0.831393	0.830636	7.569E-04	1.494E-03	-1.276E-03	3.123E-03
0.050	0.986769	0.987067	-2.982E-04	4.197E-04	4.197E-04	0
0.070*	0.999804	0.999858	-5.442E-05	1.167E-04	-4.116E-04	0
0.090**	1	1	-2.559E-07	6.902E-07	-2.936E-06	0

* 2개의 반복에서 근사값을 구하지 못함. ** 5개의 반복에서 근사값을 구하지 못함.

<표 3.3> $K=6$일 때, 정확한 검정력과 근사 검정력과 분산추정치 (변수명: FM_SIZE)

$$\hat{V}_1+\hat{V}_2=\begin{bmatrix} .00019137 & & & \\ -.00008661 & .00009771 & & \\ -.00003962 & -8.099E-06 & .00003732 & \\ -.00003689 & -4.654E-06 & 6.828E-06 & .00003223 \\ -.00001715 & 1.023E-06 & 2.231E-06 & 9.147E-07\ .00001383 \end{bmatrix}$$

norm	정확한 확률	근사 확률	정확한 확률 - 근사 확률			
			평균	표준편차	최소	최대
0	0.05	0.05	0.05	0	0.05	0.05
0.002	0.051833	0.051832	6.047E-07	1.521E-06	5.048E-08	1.278E-05
0.004	0.057598	0.057593	5.232E-06	7.951E-06	7.539E-07	3.411E-05
0.006	0.067999	0.067987	1.175E-05	1.169E-05	-3.268E-05	3.862E-05
0.008	0.083856	0.083836	1.988E-05	4.013E-05	-3.197E-05	3.915E-04
0.010	0.105670	0.105632	3.793E-05	1.419E-04	-3.127E-05	1.433E-03
0.020	0.285222	0.285002	2.196E-04	4.632E-04	-3.862E-04	1.748E-03
0.030	0.504036	0.503593	4.428E-04	6.436E-04	-7.947E-04	1.742E-03
0.040	0.707341	0.706504	8.373E-04	7.788E-04	-8.014E-04	1.748E-03
0.060	0.947276	0.947317	-4.053E-05	4.957E-04	-8.010E-04	8.607E-04
0.080	0.996846	0.997113	-2.666E-04	2.913E-04	-7.549E-04	0
0.100*	0.999947	0.999964	-1.755E-05	2.816E-05	-8.511E-05	0
0.150**	1	1	-3.934E-12	0	0	0

* 1개의 반복에서 근사값을 구하지 못함. ** 9개의 반복에서 근사값을 구하지 못함.

<표 3.4> $K=6$일 때, 정확한 검정력과 근사 검정력 및 분산추정치 (변수명: FWRKLWCT)

$$\hat{V}_1+\hat{V}_2=\begin{bmatrix} .00013506 & & & & \\ -.00005588 & .00009060 & & & \\ -.00007373 & -.00003351 & .00011230 & & \\ -5.170E-06 & -1.120E-06 & -4.000E-06 & .00001058 & \\ -6.547E-07 & 2.124E-08 & -7.137E-07 & -1.703E-07 & 1.504E-06 \end{bmatrix}$$

norm	정확한 확률	근사 확률	정확한 확률 - 근사 확률			
			평균	표준편차	최소	최대
0.000	0.05	0.05	0.05	0	0.05	0.05
0.002	0.066330	0.066323	6.775E-06	1.360E-05	-3.309E-05	3.896E-05
0.004	0.126586	0.126486	1.006E-04	3.363E-04	-3.356E-05	1.746E-03
0.006	0.210041	0.209866	1.751E-04	4.543E-04	-7.998E-04	1.696E-03
0.008	0.285243	0.285046	1.977E-04	5.444E-04	-8.019E-04	1.734E-03
0.010	0.346380	0.346243	1.367E-04	4.938E-04	-8.001E-04	1.748E-03
0.020	0.550894	0.550765	1.286E-04	5.391E-04	-7.987E-04	1.749E-03
0.030	0.713423	0.713069	3.541E-04	6.122E-04	-7.992E-04	1.747E-03
0.040	0.853332	0.852681	6.515E-04	8.455E-04	-8.009E-04	1.749E-03
0.060	0.985697	0.985969	-2.721E-04	3.080E-04	-7.928E-04	0
0.080	0.999720	0.999779	-5.851E-05	9.352E-05	-3.382E-04	0
0.100*	0.999999	1.000000	-5.216E-07	1.059E-06	-5.993E-06	0
0.150**	1	1	-1.210E-16	5.755E-16	-5.440E-15	0

* 1개 반복에서 근사값을 구하지 못함 ** 9개의 반복에서 근사값을 구하지 못함.

<표 3.5> $K=8$일 때, 정확한 검정력과 근사 검정력 및 분산추정치 (변수명: FHICOST)

$\hat{V}_1+\hat{V}_2=$

$$\begin{bmatrix} .00012195 & & & & & & \\ -.00005497 & .00009075 & & & & & \\ -.00003137 & -.00003501 & .00009704 & & & & \\ -.00001483 & 2.060E-06 & -.00001373 & .00002965 & & & \\ -8.991E-06 & 2.338E-06 & -3.756E-06 & 2.981E-07 & .00001215 & & \\ -9.332E-06 & 7.356E-06 & -.00001076 & 1.882E-06 & 1.349E-06 & .00001019 & \\ -1.249E-06 & -1.260E-06 & -3.853E-06 & -3.770E-06 & -1.774E-06 & 2.526E-07 & 9.828E-06 \end{bmatrix}$$

norm	정확한 확률	근사 확률	정확한 확률 - 근사 확률			
			평균	표준편차	최소	최대
0.000	0.05	0.05	0.05	0	0.05	0.05
0.002	0.052225	0.052225	4.201E-07	6.486E-07	5.681E-08	4.893E-06
0.005	0.064800	0.064791	8.679E-06	7.622E-06	1.942E-06	3.154E-05
0.010	0.119662	0.119640	2.243E-05	4.130E-05	-2.552E-05	4.025E-04
0.020	0.375866	0.375592	2.742E-04	4.119E-04	-5.557E-04	1.145E-03
0.030	0.670004	0.669608	3.968E-04	5.193E-04	-5.718E-04	1.145E-03
0.040	0.864498	0.864177	3.215E-04	6.622E-04	-5.718E-04	1.145E-03
0.050	0.959678	0.959720	-4.246E-05	3.223E-04	-5.718E-04	5.380E-04
0.060	0.992129	0.992349	-2.196E-04	2.449E-04	-5.719E-04	0
0.070	0.999067	0.999166	-9.868E-05	1.422E-04	-3.861E-04	0
0.080	0.999936	0.999952	-1.611E-05	2.669E-05	-8.150E-05	0
0.090	0.999998	0.999999	-1.115E-06	2.077E-06	-6.862E-06	0
0.100*	1	1	-3.483E-08	7.196E-08	7.196E-08	0

* 1개의 반복에서 근사값을 구하지 못함.

이러한 Solomon-Stephens의 3차적률카이제곱근사법은 비교적 참 분포에 근사한 결과를 보여줄 뿐 아니라 컴퓨터의 발달과 더불어 실용성이 더욱 높아졌다. 본 연구에서 우리는 Solomon-Stephens의 3차적률카이제곱근사법의 구체적인 도출과정을 제시하고, 동질성검정을 위한 Wald 검정의 검정력을 통해 비중심모수가 0이 아닌 경우에도 그들의 3차적률카이제곱근사법이 참값에 매우 근사한 결과를 제공함을 확인할 수 있었다.

참고문헌

[1] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics,* Vol. 25, 290-302.

[2] Fraser, D.A.S., Wang, A.C.M. and Wu. J. (1998). An approximation for the non-central chi-squared distribution. *Communications in Statistics-Simulation,* Vol. 27(2), 275-287.

[3] Graybill, F.A. (1976). *Theory and Application of Linear Model.* Belmont, CA: Wadsworth.

[4] Gurland, J. (1953). Distribution of quadratic forms and ratios of quadratic forms. *The Annals of Mathematical Statistics,* Vol. 24, 416-427.

[5] Gurland J. (1954). Distribution of definite and indefinite quadratic forms. *The Annals of Mathematical Statistics,* Vol. 26, 122-127.

[6] Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, Vol. 48, 419-426.

[7] Jensen, D.R. and Solomon, H. (1972). A Gaussian approximation to the distribution of a definite quadratic form. *Journal of American Statistical Association,* Vol. 67, 898-902

[8] Kerridge, D. (1965). A probabilistic derivation of the non-central χ^2distribution. *The Australian Journal of Statistics,* Vol. 7, 37-39.

[9] Pearson, E.S. (1959). Note on an approximation to the distribution of noncontral χ^2distribution. *Biometrika*, Vol. 44, 364.

[10] Posten, H.O. (1989). An effective algorithm for the noncentral chi-squared distribution function. *The American Statistician,* Vol. 43, 261-263.

[11] Rao, J.N.K. and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit the independence in two-way tables. *Journal of the American Statistical Association,* Vol. 76, 221-230.

[12] Solomon, H. and Stephens, M.A. (1977). Distribution of a Sum of Weighted Chi-Square Variables. *Journal of the American Statistical Association,* Vol. 72, 881-885.

[13] Thisted, R.A. (1988). *Elements of Statistical Computing.* Chapman and Hall/CRC, New York.

[14] Tiku, M. (1985), Noncentral chi-square distribution. *Encyclopedia of Statistical Sciences* (9 Vols. plus Supplement), Vol. 6, 276-280

[15] U.S. National Center for Health Statistics (2004). *Data File Documentation, U.S. National Health InterviewSurvey, 2003* (machine readable data file and documentation). U. S. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland, U. S.

[Received May 2006, Accepted July 2006]