비중심카이제곱분포 확률계산의 비교

구 선 희†

요 약

비중심 χ^2분포의 누적분포 함수의 계산은 χ^2검정에서 검정력 계산에 요구된다. 비중심 χ^2분포의 확률계산은 일반적으로 많은 시간이 소요될 뿐 아니라 정확성이 문제가 되어 여러 접근방법들이 계산속도와 정확성을 고려하여 제시되었다. 따라서 본 논문에서는 기존의 비중심 χ^2분포의 확률 계산 접근방법에 의한 계산결과를 비교하고 자유도와 비중심모수에 따른 효율적인 접근방법을 탐색하였다.

Accuracy and Efficiency for Computation of Noncentral χ^2 Probabilities

Son Hee Gu†

ABSTRACT

The evaluation of the cumulative distribution function of the noncentral χ^2 distribution is required in approximate determination of the power of the χ^2 test.

Many approximations to the cumulative distribution function of the noncentral χ^2 distribution have been suggested. However, in selecting an approximations both simplicity and accuracy should be considered.

In this note we compared various approximations in terms of accuracy and efficiency.

1. 서 론

비중심 χ^2분포는 1928년 R. A. Fisher에 의해 다중상관계수 R의 분포를 계산하는 과정에서 유도 되었으며[6], Miller와 Park 등에 의하면 비중심 χ^2분포는 일반화된 Rayleigh 분포로서 Rayleigh-Rice 또는 Rice 분포라고 불리우고 있다[9, 10]. 이와같은 비중심 χ^2분포의 확률 계산을 통하여 범주형자료의 분석에 흔히 사용되는 적합도검정 및 독립성, 동질성검정 등과 같은 χ^2검정의 검정력을 구할 수 있다[4, 8]. 비중심 χ^2분포에서 분포함수(distribution function)의 값을 계산하는데에는 일반적으로 많은 시간이 소요될 뿐 아니라 정확성이 문제가 되고 있다. 이로 인하여 비중심 χ^2분포의 확률 계산에 대한 여러 접근방법들이 제시되었다. 따라서 본 논문에서는 비중심 χ^2분포의 확률 계산에 있어 계산속도와 정확성을 고려하여 효율적인 방법을 모색하고자 한다. 이를 위해 기존의 비중심 χ^2분포의 확률 계산 접근방법에 의한 계산결과를 비교하고 자유도와 비중심모수에 따른 효율적인 접근방법을 탐색하였다.

2. 비중심 χ^2분포의 확률 계산 접근방법

Fisher는 Bessel함수로서 비중심 χ^2분포의 누적함수를 정의하였으며[6], 분포식으로부터 일부 정확한 확률값을 제시하였다. 그러나 이 분포의 확률 계산을 하기 위해서는 상당한 어려움이 수반된다. 따라서 비중

† 정 회 원 : 전주대학교 전기전자컴퓨터 공학부 재원교수
논문접수 : 1996년 5월 25일, 심사완료 : 1996년 11월 20일

심 χ^2분포의 확률 계산에 빠르고 정확한 확률값을 구하기 위하여 다음과 같은 접근방법들이 제시되었다.

2.1 Abdel-Aty의 방법

Abdel-Aty는 비중심 χ^2분포를 정규화하는 접근방법으로 '초기 접근방법'을 제시하였다[1].

비중심 χ^2분포의 확률밀도함수는

$$p_{\chi_v'^2}(x)=\sum_{j=0}^{\infty}\left[\frac{(\frac{1}{2}\lambda)^j}{j!}e^{-\frac{1}{2}\lambda}\right]p_{\chi^2_{2j+v}}(x) \qquad (2.1)$$

이다. 여기서 v는 자유도, λ는 비중심모수 이다. 또한 s차 누가적률(cumulant moment)은

$$\kappa_s=2^{s-1}(s-1)!(v+s\lambda)$$

이다. 여기에서는 비중심 χ^2분포를 정규화하는 접근방법으로 $\kappa_2(y)$에서 $(v+\lambda)^2$의 계수를 최소화하고 $s>2$에 대하여 $\kappa_s(y)$를 최소화하는 것으로 다음 식을 고려한다.

$$y=\left(\frac{\chi'^2}{v+\lambda}\right)^{\frac{1}{3}}$$

y의 누가적률로부터 다음식이 유도된다.

$$\left(\frac{\chi'^2}{v+\lambda}\right)^{\frac{1}{3}}\sim N\left(1-\frac{2}{9}\frac{(v+2\lambda)}{(v+\lambda)^2},\ \frac{2}{9}\frac{(v+2\lambda)}{(v+\lambda)^2}\right)$$

2.2 Sankaran의 방법

Sankaran은 비중심 χ^2분포를 정규화하는 접근방법으로 다음 식을 고려하였다[13].

$$z=\left[\frac{\chi'^2-\frac{1}{2}(v-1)}{v+\lambda}\right]^{\frac{1}{2}}$$

z의 누가적률로부터 다음식이 유도된다.

$$\left\{\chi'^2-\frac{1}{2}(v-1)\right\}^{\frac{1}{2}}\sim N\left\{\left(\lambda+\frac{1}{2}(v-1)\right)^{\frac{1}{2}},1\right\}$$

2.3 Arthur와 Chou의 방법

Arthur와 Chou는 비중심 χ^2분포를 유도하기 위하여 $\chi_v'^2(\lambda)=\chi_1'^2(\lambda)+\chi^2_{v-1}$의 관계식을 이용하였다[3].

즉, 자유도가 1이고 비중심모수가 λ인 비중심 χ^2분포와 자유도가 $v-1$인 중심 χ^2분포로부터 $\chi_v'^2(\lambda)$의 누적분포함수는 다음과 같다.

$$v=1:\Pr[\chi_1'^2(\lambda)\le u]$$
$$=G(\sqrt{u}-\sqrt{\lambda})-G(-\sqrt{u}-\sqrt{\lambda})$$
$$v\ge 2: F(u;v,\lambda)=\Pr[\chi_v'^2(\lambda)\le u]$$
$$=\frac{\sqrt{2\pi}}{\Gamma\left(\frac{v-1}{2}\right)2^{\frac{v-1}{2}}}$$
$$\times\int_0^u y^{\frac{v-3}{2}}G'(\sqrt{y})[G(\sqrt{u-y}-\sqrt{\lambda})$$
$$-G(-\sqrt{u-y}-\sqrt{\lambda})]\,d\lambda$$

여기서 $G(z)=\int_{-\infty}^{z}G'(t)\,dt$,

$G'(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ 이다.

2.4 Posten의 방법

Posten은 비중심 χ^2분포가 포아송분포와 중심 χ^2분포 형태를 통하여 제시되는 특성을 이용하였다[12]. 즉, 중심 χ^2분포에 순환적 알고리즘을 적용하여 계산한 확률을 다시 비중심 χ^2분포에 순환적 알고리즘을 적용하는 접근방법을 제시하였다.

식(2.1)과 같이 비중심 χ^2분포는 포아송분포 $P(j)$와 중심 χ^2분포 $W(j)$의 곱으로 나타낼 수 있다[7].

$$\Pr(\chi_v'^2(\lambda)\le x)=\sum_{j=0}^{\infty}P(j)W(j)$$

여기서 $P(j)=\frac{\left(\frac{\lambda}{2}\right)^j e^{-\frac{\lambda}{2}}}{j!}$,

$W(j)=\Pr[\chi^2_{v+2j}\le x]$ 이다.

이로 부터 다음 식이 유도된다.

$$W(j)=-S(j)+W(j-1)$$

여기서 $S(j)=\frac{x^{\frac{v+2j-2}{2}}e^{-\frac{x}{2}}}{\Gamma\left(\frac{v+2j}{2}\right)2^{\frac{v+2j-2}{2}}}\quad j=0,1,2,\cdots$

이다.

또한 $S(j)=\dfrac{x}{v+2j-2}\ S(j-1),\ j=1, 2, \cdots$ 와

$$S(0)=\frac{x^{\frac{v-2}{2}}\ e^{-\frac{x}{2}}}{\Gamma\left(\frac{v}{2}\right)\ 2^{\frac{v-2}{2}}}$$

그리고 $W(0)=\Pr[\chi_v^2\le x]$로부터 $W(j)$를 계산한다.

$p(0)=e^{-\frac{\lambda}{2}}$ 와

$$P(j)=\left(\frac{\lambda}{2j}\right)\ P(j-1), \qquad j=1, 2, \cdots$$

으로 부터

$$\Pr(\chi_v'^2(\lambda)\le x)=\sum_{j=0}^{\infty} P(j)W(j)$$

을 구할 수 있다.

$W(0)=\Pr[\chi_v^2\le x]$의 계산은 Boardman과 Kopitzke 그리고 Posten 등이 사용한 연분수(continued fraction)방법을 적용하였다[5, 11].

2.5 Abdel-Samad와 Ashour의 방법

Abdel-Samad와 Ashour는 Shea가 제시한 중심 χ^2분포의 확률 계산 접근방법을 이용하여 비중심 χ^2분포의 확률 계산 접근방법을 제시하였다[2, 14].

식(2.1)과 같이 Shea가 제시한 중심 χ^2분포 함수를 적용한 비중심 χ^2분포 함수는 다음과 같다.

$$F(u;v,\lambda)=e^{-\frac{\lambda}{2}}\ f_v(x)\times\sum_{j=0}^{\infty}\frac{1}{j!}\ C_j\left(\frac{\lambda x}{4},\frac{v}{2}\right)$$

$$\times\left\{\sum_{s=0}^{\infty} C_s\left(\frac{x}{2},\frac{v}{2}+j\right)\right\}$$

여기서 $P(j)=\dfrac{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}+1\right)}$ 이며

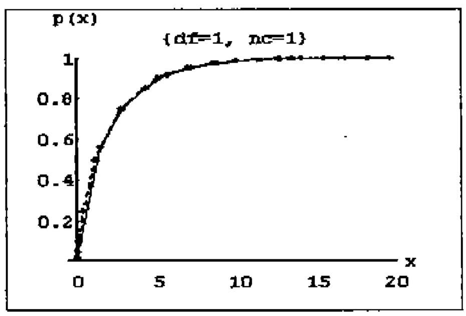
$$C_j\left(\frac{\lambda x}{4},\frac{v}{2}\right)=\frac{\frac{\lambda x}{4}}{\frac{v}{2}+j}\ C_{j-1}\left(\frac{\lambda x}{4},\frac{v}{2}\right),\ j=1, 2, 3, \cdots$$

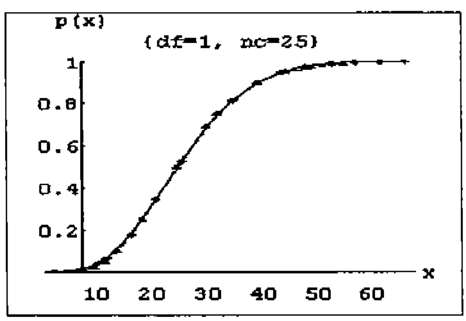
$$C_0\left(\frac{\lambda x}{4},\frac{v}{2}\right)=1$$

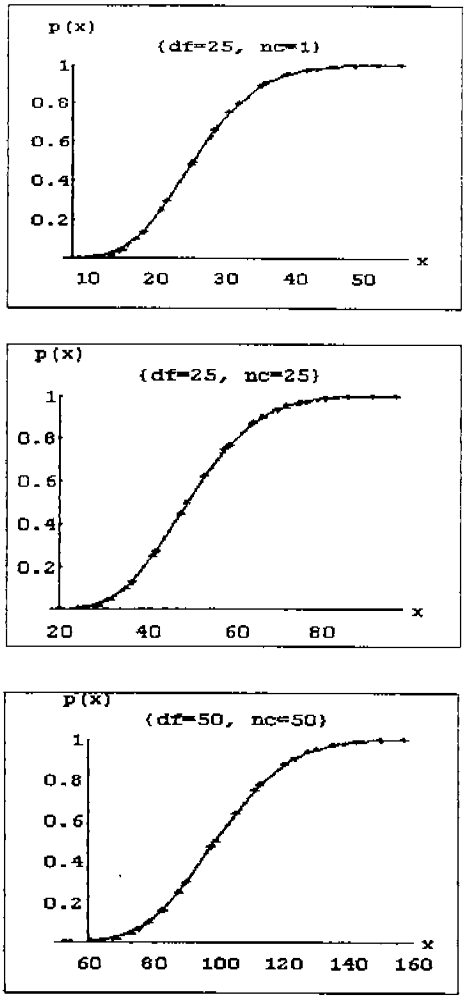
이다.

3. 확률 계산의 비교

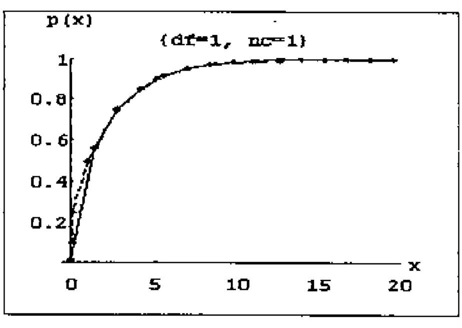
다음은 일부 자유도와 비중심모수에 대하여 Fisher의 참값과 각 각의 접근방법에 의한 근사값과의 차이를 그림을 통하여 살펴보았다. 접근방법의 특성에 따라 다음과 같이 (그림 1, 2)과 (그림 3, 4, 5)로 나누어서 비교하였다. (그림 1, 2)는 자유도, 비중심모수, 확률이 주어진 경우에 백 분위수를 계산하였으며, (그림 3, 4, 5)는 자유도, 비중심모수, 백 분위수가 주어진 경우에 확률을 계산하였다. 여기에서 제시하고 있는 확률값은 PC-486에 Mathematica Program을 이용하여 계산하였다.

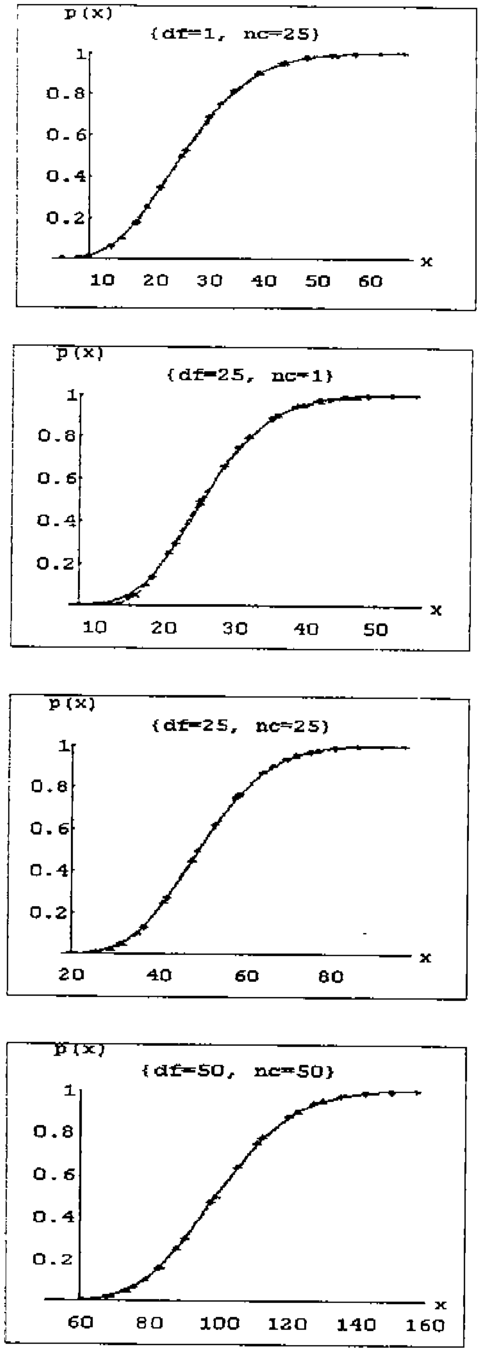






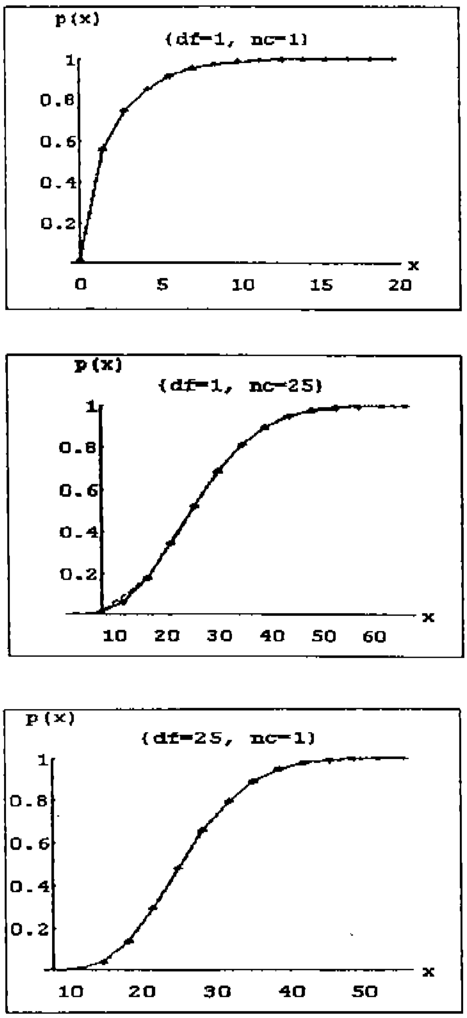
(그림 1) Fisher의 참값과 Abdel-Aty의 근사값 비교
*df : 자유도, nc : 비중심모수
*----- : Fisher의 확률값
- - - : Abdel-Aty의 확률값
(Fig. 1) Comparison of the Fisher's value and Abdel-Aty's value

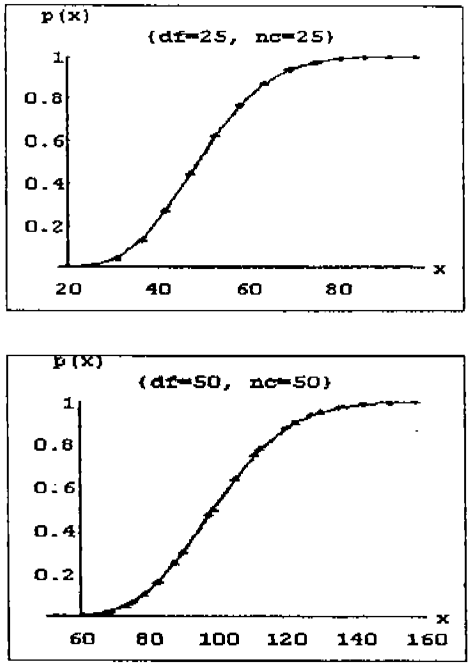




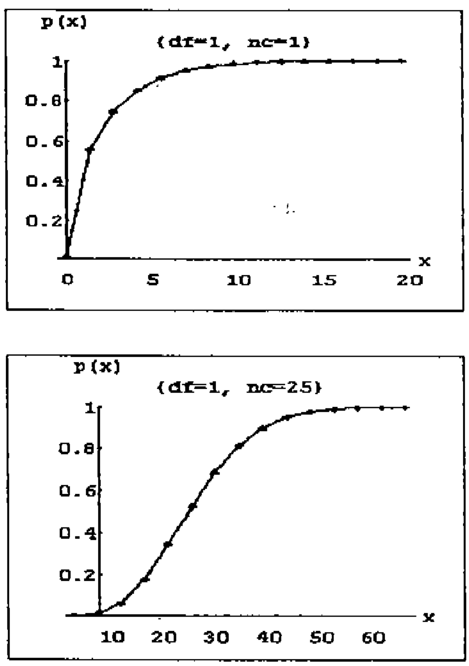
(그림 2) Fisher의 참값과 Sankaran의 근사값 비교
*df : 자유도, nc : 비중심모수
*----- : Fisher의 확률값
- - - : Sankaran의 확률값
(Fig. 2) Comparison of the Fisher's value and Sankaran's value

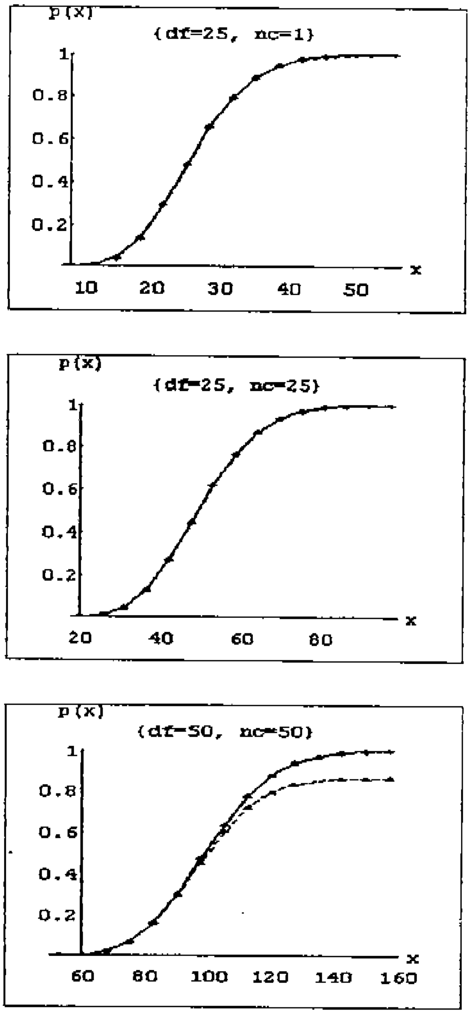
그림으로부터 다음과 같은 사실을 알 수 있다. 다섯가지 접근방법중 초기의 접근방법은 정규분포로 정규화하여 확률 계산을 하였는데 Abdel-Aty의 접근방법은 자유도가 클수록, Sankaran의 접근방법은 비중심모수가 클수록 Fisher의 비중심 χ^2분포의 참값과 근접하게 나타나고 있다. 또한 두 접근방법 모두 자유도와 비중심모수가 30 이상일 때는 Fisher의 참값과 상당히 근접하게 나타나고있음을 알 수 있다. 여기서 Abdel-Aty와 Sankaran의 접근방법을 통하여 하나의 확률값을 구하기 위하여 소요된 시간은 두 접근방법 모두 0.055초 이며 Fisher의 확률값을 구하는 데 소요되는 시간 4.6초와 비교하면 상당히 계산속도가 향상됨을 알 수 있다.



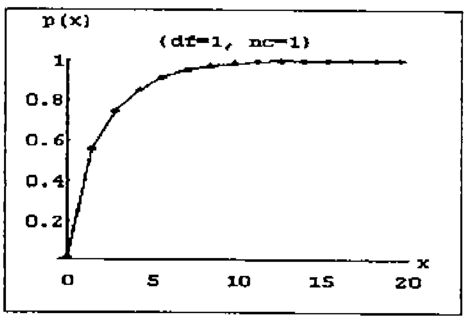


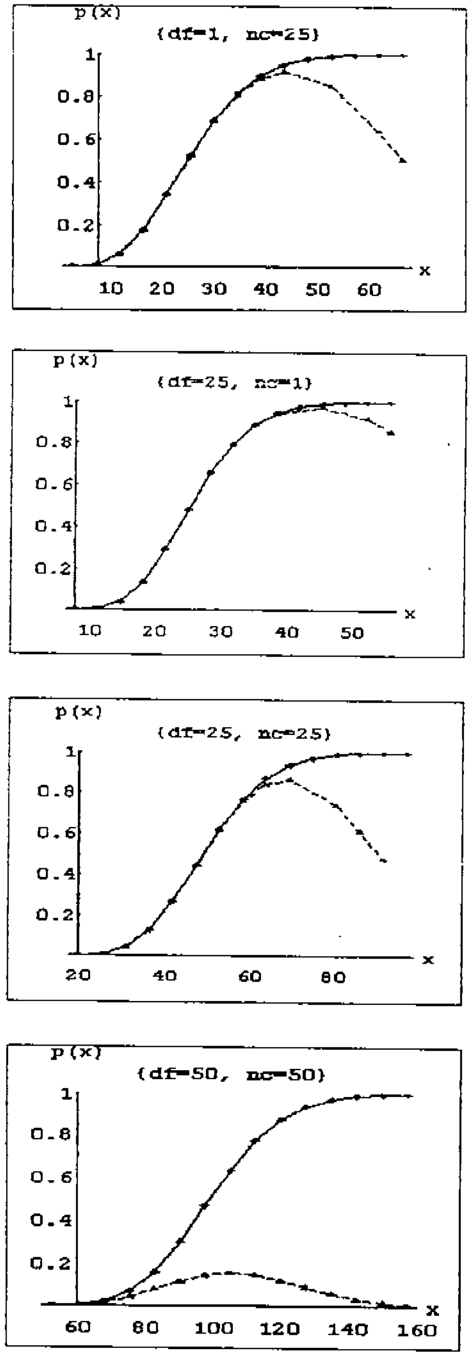
(그림 3) Fisher의 참값과 Arthur와 Chou의 근사값 비교
* df : 자유도, nc : 비중심모수
* ----- : Fisher의 확률값
- - - : Arthur와 Chou의 확률값
(Fig. 3) Comparison of the Fisher's value and Arthur and Chou's value





(그림 4) Fisher의 참값과 Posten의 근사 값 비교
*df : 자유도, nc : 비중심모수
*----- : Fisher의 확률값
- - - : Posten의 확률값
(Fig. 4) Comparison of the Fisher's value and Posten's value





(그림 5) Fisher의 참값과 Abdel-Samad와 Ashour의 근사 값 비교
*df : 자유도, nc : 비중심모수
*----- : Fisher의 확률값
- - - : Abdel-Samad와 Ashour의 확률값
(Fig. 5) Comparison of the Fisher's value and Abdel-Samad and Ashour's value

1980년 이후부터는 비중심 χ^2분포의 분포함수를 유도하는 직접적인 접근방법들이 제시되었다. Arthur와 Chou 그리고 Posten이 제시한 접근방법 모두 전반적으로 Fisher의 확률값과 근사하나 Posten의 경우에는 자유도와 비중심모수가 클수록 큰 차이를 보이고 있다. 하나의 확률값 계산에 소요되는 시간은 Arthur와 Chou의 경우는 2.83초 이며 Posten의 경우는 0.0035초이다. Abdel- Samad와 Ashour의 접근방법은 자유도와 비중심모수가 작으며 확률이 0.8 이하인 경우에는 근사하나 그 밖에는 상당한 차이를 보이고 있으며 이러한 현상은 자유도와 비중심모수가 클수록 더욱 심하게 나타나고 있다. 이 방법을 적용하는 경우 소요되는 시간은 1.36초 이다.

5가지 접근방법들을 Fisher의 확률값과 비교하여 정확성의 측면에서 살펴보면 다음과 같다.

〈표 1〉 Fisher의 확률값과 접근방법들의 근사값 비교
〈Table 1〉 Comparison of the Fisher's value and Approximate values

비중심모수 \ 자유도	1	25	50
1	1(△)2(△) 3(○)4(○) 5(○)	1(○)2(△) 3(○)4(○) 5(×)	
25	1(△)2(○) 3(△)4(○) 5(×)	1(○)2(○) 3(○)4(○) 5(×)	
50			1(○)2(○) 3(○)4(×) 5(×)

* ○ : Fisher 확률값과 차이가 거의없다.
  △ : Fisher 확률값과 차이가 약간 있다.
  × : Fisher 확률값과 차이가 있다.
* 1 : Abdel-Aty의 접근방법
  2 : Sankaran의 접근방법
  3 : Arthur와 Chou의 접근방법
  4 : Posten의 접근방법
  5 : Abdel-Samad와 Ashour의 접근방법

앞의 표에서 제시한 정확성과 계산속도를 고려한 확률 계산 결과 다음과 같은 사실을 알 수 있다.

비중심카이제곱분포의 확률계산에 있어 모든 자유도와 비중심모수에 대하여 Arthur와 Chou가 제시한 접근방법이 Fisher의 확률값과 근사하게 나타났으나 다른 접근방법과 비교할 때 계산속도가 더 소요됨에 따라 효율적이지 못하다. 그러므로 정확성 및 계산속도를 고려할 때 자유도와 비중심모수가 30이하인 경우에는 Posten의 접근방법을 사용하고 자유도와 비중심모수가 30이상인 경우에는 Abdel-Aty와 Sankaran이 제시한 접근방법을 사용하는 것이 효율적이라 할 수 있다. 정확성과 계산속도를 고려하여 효율적인 측면에서 살펴보면 다음과 같다.

〈표 2〉 접근방법들의 효율성 비교
〈Table 2〉 Efficiency comparison of Approximate values

비중심모수 \ 자유도	1 ~ 30	30 이상
1 ~ 30	Posten	Abdel-Aty
30 이상	Sankaran	Abdel-Aty Sankaran

4. 결 론

본 논문은 비중심 χ^2분포의 확률 계산을 하는데 있어 정확성과 계산속도를 고려하여 효율적인 방법을 찾고자 하였다. 기존의 다섯가지 접근방법에 대하여 PC-486에 Mathematica Program을 이용하여 정확성과 계산속도를 고려한 확률 계산 결과 다음과 같은 결론을 내릴 수 있다. 정확성의 측면에서 보면 Arthur와 Chou가 제시한 접근방법만이 Fisher의 참값과 가장 근접하게 나타나고 있으나, 계산속도를 고려한 결과 자유도와 비중심모수가 30 이하인 경우에는 Posten이 제시한 접근방법을 사용하고 자유도와 비중심모수가 30 이상인 경우에는 Abdel-Aty와 Sankaran이 제시한 접근방법을 사용하는 것이 효율적이라는 결론을 내릴 수 있다.

참 고 문 헌

[1] S. H. Abdel-Aty, "Approximate formule for the percentage points and the probability integral of the non-central χ^2-distribution," Biometrika, Vol. 41, pp.538-540, 1954.

[2] A. I. Abdel-Samad and S. K. Ashour, "On the computation of non-central chi-square distribution function," Communications in Statistics, Simulation, Vol.19(4), pp.1279-1291, 1990.

[3] K. H. Arthur and Y. M. Chou, "New representations of the noncentral chi-square density and cumulative," Communications in Statistics, Theory and Methods, Vol.A13(21), pp.2673-2678, 1984.

[4] P. J. Bickel and K. A. Doksum, 'Mathematical Statistics: Basic ideas and selected topics,' San Francisco: Holden-Day, 1977.

[5] T. J. Boardman and R.W.Kopitzke, "Probability and table values for statistical distributions," in Proceedings of the Statistical Computing Section, American Statistical Association, pp.81-86, 1975.

[6] R. A. Fisher, "The general sampling distribution of the multiple correlation coefficient," Proceedings of the Royal Society of London, Vol.121A, pp.654-673, 1928.

[7] N. L. Johnson, "On an extension of the connexion between Poisson and χ^2-distributions," Biometrika, Vol.46, pp.352-363, 1959.

[8] N. L. Johnson and S.Kotz, 'Continuous Univariate Distributions,' Boston: Houghton Mifflin, 1970.

[9] K. S. Miller and R. I. Bernstein and L. E. Blumenson, "Generalized Rayleigh Processes," Quarterly of Applied Mathematics, Vol.16, pp.137-145, 1958.

[10] J. H. Park, "Moments of the generalized Rayleigh distribution," Quarterly of Applied Mathematics, Vol.19, pp.45-49, 1961.

[11] H. O. Posten, "Computer algorithms for the classical distribution functions," in Proceedings for the First International Conference on the Teaching of Statistics, Sheffield, U, K.: University of Sheffield Printing Unit, pp.313-330, 1982.

[12] H. O. Posten, "An effective algorithm for the noncentral chi-squared distribution function," The American Statistician, Vol.43, pp.261-263, 1989.

[13] M. Sankaran, "Approximations to the non-central chi-square distribution," Biometrika, Vol.50, pp.199-204, 1963.

[14] B. L. Shea, "Chi-squared and incomplete Gamma integral," Applied Statistics, Statistical Algorithms, Vol.37, pp.466-473, 1988.

구 선 희

1986년 숙명여자대학교 수학과 졸업(이학사)

1989년 숙명여자대학교 대학원 수학과 졸업(이학석사)

1996년 성균관대학교 대학원 통계학과 졸업(경제학박사)

1996년~현재 전주대학교 전기전자컴퓨터 공학부 객원교수

관심분야: 신경망이론, 전산수학