Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No. 1, pp. 83 ~ 90

주변값이 주어진 이원분할표에 대한 카이제곱 검정통계량의 소표본 분포 및 대표본 분포와의 일치성 연구

박철용[1] · 최재성[2] · 김용곤[3]

요 약

이원분할표의 두 범주형 변수에 대한 독립성을 검정할 때 흔히 카이제곱 검정통계량이 사용된다. 표본추출 모형이 다항이나 곱다항인 경우 이 검정통계량이 독립성 가정하에서 근사적으로 카이제곱 분포를 따르게 되는 것은 잘 알려진 사실이다. 두 주변값이 모두 주어진 경우 독립성 가정하에서 표본추출 모형은 다중 초기하분포가 되며 앞의 모형과 마찬가지로 카이제곱 통계량에 근거한 검정을 사용할 수 있다.

이 연구에서는 주변값이 주어진 경우에 카이제곱 통계량의 소표본 분포를 대표본 분포인 카이제곱 분포와 비교하고자 한다. 표본크기가 작은 몇개의 경우에 대해 카이제곱 통계량의 소표본 분포를 직접 계산해보았다. 표본크기가 큰 몇개의 경우는 간단한 몬테칼로 알고리듬을 통해 소표본 분포를 생성하고 카이제곱 확률도와 콜모고로브-스미노브 단일표본 검정을 이용하여 대표본 분포와의 일치성을 알아보았다.

주제어: 다중 초기하분포, 독립성 검정, 몬테칼로 알고리듬, 콜모고로브-스미노브 단일표본 검정

1. 서론

관찰도수가 $\{n_{ij}\}$인 이원분할표를 고려해 보자. 이 때 행과 열에 대응되는 범주형 변수들이 서로 독립이라는 가설을 검정하는 카이제곱 통계량은 다음과 같이 주어진다.

$$X = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} \tag{1}$$

여기서 $n_{i+} = \sum_{j=1}^{c} n_{ij}$는 행의 주변값들(margins), $n_{+j} = \sum_{i=1}^{r} n_{ij}$는 열의 주변값들이며 $n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$은 표본크기이다. 위의 검정통계량은 표본추출 모형이 다항모형(multinomial

[1](704-701) 대구 달서구 신당동 1000번지, 계명대학교 통계학과 부교수
[2](704-701) 대구 달서구 신당동 1000번지, 계명대학교 통계학과 교수
[3](704-701) 대구 달서구 신당동 1000번지, 계명대학교 통계학과 조교수

model), 곱다항모형(product-multinomial model)인 경우에 모두 적용되며 표본크기가 커지게 되면 근사적으로 $\chi^2((r-1)(c-1))$ 분포를 따르게 된다. 그런데 곱다항모형은 다항모형에서 행(혹은 열)의 주변값들이 주어진 조건부 분포 형태로 나타나기 때문에 직감적으로 앞의 결과가 행과 열 모두의 주변값들이 주어진 모형에서도 적용될 수 있으리라 생각해볼 수 있다. (구체적인 증명은 Alalouf (1987)와 Park (1995) 등에서 찾아볼 수 있다.) 두 주변값이 모두 주어진 모형에서는 독립성 가정 하에서 관찰도수의 분포가 다중 초기하분포(multiple hypergeometric distribution)를 따르게 되며 그 구체적인 모양은 다음과 같이 주어진다.

$$P(\{n_{ij}\} \mid \{n_{i+}, n_{+j}\}) = \frac{\prod_{j=1}^{c} \binom{n_{+j}}{n_{1j},\ldots,n_{rj}}}{\binom{n}{n_{1+},\ldots,n_{r+}}} = \frac{(\prod_{i=1}^{r} n_{i+}!)\left(\prod_{j=1}^{c} n_{+j}!\right)}{n! \prod_{i=1}^{r}\prod_{j=1}^{c} n_{ij}!} \tag{2}$$

주어진 이원분할표의 독립성 가정에 대해 위의 소표본 분포에 근거하여 검정을 행하는 방법이 Freeman과 Halton (1951)의 정확검정으로서 Fisher의 2×2분할표에 대한 정확검정을 연장한 것이다. 여기서 p값(p-value)은 관측된 분할표와 발생확률이 같거나 작은 모든 분할표들의 확률을 합한 것이 되는데 Mehta와 Patel (1983)의 네트워크 알고리듬(network algorithm)의 도움으로 상용 통계팩키지인 SAS의 FREQ 절차에서도 이 값을 계산할 수 있다. 그런데 (1)의 검정통계량에 근거한 정확검정의 p값은 관측된 분할표의 통계량값과 같거나 큰 통계량값을 가지는 모든 분할표들의 확률을 합한 것이 된다. 따라서 (1)의 검정통계량의 값이 (2)에 근거한 분할표 발생확률의 단조함수가 아니기 때문에 (1)의 검정통계량에 근거한 정확검정과 Freeman과 Halton의 정확검정은 다른 결과를 제시한다. (1)과 같이 특정 검정 통계량에 근거한 정확검정을 수행하기 위해서는 윈도우즈 버전 6.04의 SAS FREQ와 정확한 비모수 추론을 위한 특수목적 소프트웨어인 StatXact (1991)를 사용할 수 있다.

이 논문에서는 (1)에서 주어진 검정통계량의 소표본 분포가 어느 정도 대표본 분포인 $\chi^2((r-1)(c-1))$에 가까운지 알아보고자 한다. 그런데 검정통계량의 정확분포를 계산하기 위해서는 엄청난 계산시간이 필요로 한다. 주어진 분할표에 대한 카이제곱 정확검정의 경우 네트워크 알고리듬의 도움으로 계산과정이 상당히 단축되지만, 검정통계량의 정확분포를 계산하기 위해서는 모든가능한 분할표에 대해 검정통계량의 값을 계산하여야 하기 때문에 표본크기가 조금만 커지게 되어도 엄청난 계산시간을 필요로 한다. 따라서 이 논문에서는 간단한 몬테칼로 알고리듬을 통해 소표본 분포를 생성하고 카이제곱 확률도(chi-squared probability plot)와 Kolmogorov-Smirnov 단일표본 검정을 통해 이 소표본 분포가 대표본 분포와 얼마나 가까운지 판단하고자 한다.

논문의 나머지 부분은 다음과 같이 구성되었다. 2절에서는 표본크기가 작은 몇개의 경우에 대해 카이제곱 통계량의 정확분포를 계산해보고, 표본크기가 큰 몇개의 경우에 대해 몬테칼로 알고리듬 방법을 통해 소표본 분포를 생성하고 대표본 분포와 얼마나 가까운지 알아본다. 3절에서는 2절에서 얻은 결과를 바탕으로 전체적인 결론을 내린다.

2. 카이제곱 통계량의 소표본 분포 및 대표본 분포와의 일치성

우선 표본크기가 8 이하인 대표적인 분할표들에 대해 카이제곱 통계량의 정확분포를 계산해보도록 하자. 여기서는 계산상의 편의를 위해 카이제곱 통계량의 정확분포를 바로 제시하지 않고 가능한 분할표에 따른 (1)의 카이제곱 통계량값과 그에 해당되는 (2)의 확률값을 제시한다. 또한 표기상의 편의를 위해 행렬 $\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$ 을 간단히 $(n_{11}\ n_{12};n_{21}\ n_{22})$ 형태로 나타내기로 한다. 표본크기가 8 이하인 대표적인 분할표에 대한 결과가 <표 1>에 제시되어 있다.

<표 1>에서 알 수 있듯이 행과 열의 주변값이 (1,1)과 (1,1)으로 되어 있는 경우 극단적으로 카이제곱 검정통계량이 하나의 값만 취하며 나머지 경우에도 많아야 네 개의 값만 취한다. 여기에서 우리가 주목해야 하는 사실은 행과 열의 주변값들이 동질적일수록 분할표의 개수는 증가한다는 것이다. Gail과 Mantel (1977)은 가능한 모든 분할표의 개수를 계산해주는 돠부름 공식(recursion formula)과 근사공식을 제시하였는데, 가능 분할표 개수는 표본크기가 커지면 기하급수적으로 증가하며, 또한 분할표의 크기와 주변값의 동질성이 커지면 빠르게 증가하게 되는 것으로 알려져 있다. 예를들어 동일한 주변값을 가진 분할표의 총가능 개수가 표본크기가 20일 때는 40,176인데 표본크기가 100이 되면 근사적으로 8.4×10^9이나 되는 것이다.

위에서 살펴본 바와 같이 표본크기가 조금만 커지게 되면 생기는 엄청난 계산시간의 문제점을 극복하기 위해서 다음과 같이 간단한 몬테칼로 알고리듬을 이용하여 카이제곱 통계량의 소표본 분포를 생성하고자 한다.

- 단계 1: 주어진 행의 주변값 만큼 발생빈도가 있는 원자료 행 범주값 벡터 x 를 만든다. 마찬가지 방법으로 원자료 열 범주형 벡터 y를 만든다.
- 단계 2: 열 범주형 벡터 y를 무작위 재배열하여 y^*를 생성한다.
- 단계 3: x와 y^*에 근거하여 이원분할표를 구성하여 (1)의 검정통계량값을 계산한다.
- 단계 4: 앞의 1-3 단계를 여러번 되풀이한다.

앞의 1-3 단계에 의해 (2)와 같은 확률을 가지는 이원분할표가 생성된다는 것은 직감적으로 쉽게 알 수 있다. (구체적인 논증이 필요하면 허명회 (1997)를 참조할 수 있다.) 이와 다른 모의실험 방법들이 (예를들어 Patefiled (1981) 참조) 제시되고 있지만 이 논문에서는 이해하기 쉽고 논증이 쉬운 앞의 알고리듬을 따르기로 한다. 이 논문에서 시행하는 모의실험에서는 1000번의 반복계산을 통해 카이제곱 통계량값을 생성하고 이 값들을 대표본 분포인 $\chi^2((r-1)(c-1))$과 대비하기 위해서 두 가지 방법을 사용하였다. 첫번째 사용된 방법은 그림에 의한 비공식적인 방법인 카이제곱 확률도(chi-squared probability plot)이며 두번째 방법은 공식적인 방법인 Kolmogorov-Smirnov(K-S) 단일표본 검정이다.

표 1: 2×2 분할표에 대한 카이제곱 통계량값과 발생 확률값

행과 열의 주변값	분할표	통계량값	확률값
(1,1), (1,1)	(0 1; 1 0)	2	0.5
	(1 0; 0 1)	2	0.5
(1,2), (1,2)	(0 1; 1 1)	0.75	2/3
	(1 0; 0 2)	3	1/3
(2,2), (2,2)	(0 2; 2 0)	4	1/6
	(1 1; 1 1)	0	2/3
	(2 0; 0 2)	4	1/6
(1,3), (1,3)	(0 1; 1 2)	0.444	3/4
	(1 0; 0 3)	4	1/4
(3,3), (3,3)	(0 3; 3 0)	6	0.05
	(1 2; 2 1)	0.667	0.45
	(2 1; 1 2)	0.667	0.45
	(3 0; 0 3)	6	0.05
(2,4), (2,4)	(0 2; 2 2)	1.5	0.4
	(1 1; 1 3)	0.375	0.533
	(2 0; 0 4)	6	0.067
(1,5), (1,5)	(0 1; 1 4)	0.24	0.833
	(1 0; 0 5)	6	0.167
(4,4), (4,4)	(0 4; 4 0)	8	0.014
	(1 3; 3 1)	2	0.229
	(2 2; 2 2)	0	0.514
	(3 1; 1 3)	2	0.229
	(4 0; 0 4)	8	0.014
(3,5), (3,5)	(0 3; 3 2)	2.88	0.179
	(1 2; 2 3)	0.036	0.536
	(2 1; 1 4)	1.742	0.268
	(3 0; 0 5)	8	0.018
(2,6), (2,6)	(0 2; 2 4)	0.889	0.536
	(1 1; 1 5)	0.889	0.429
	(2 0; 0 6)	8	0.036
(1,7), (1,7)	(0 1; 1 6)	0.163	0.875
	(1 0; 0 7)	8	0.125

구체적인 모의실험은 다음과 같이 구성되었다.

첫째, 5×5분할표와 10×10분할표를 대상으로 동질적인 주변값을 가지는 경우만 대상으로 한다.

둘째, 각 차원의 분할표에 대한 표본크기는 기대도수가 1, 2, 4인 세 가지 경우를 대상으로 한다.

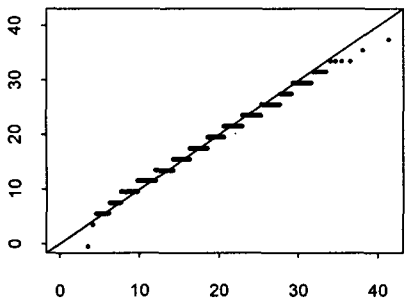
앞의 몬테칼로 알고리듬에 의한 모의실험 결과는 <그림 1>에 나타나 있다. 각 그림의 x축은 기대분위수, y축은 순서통계량이며 소표본 분포와 대표본 분포와의 일치도를 쉽게 알아볼 수 있도록 원점을 지나는 기울기가 1인 기대직선을 그려 넣었다. 각 그림 아래에는 공식적인 방법인 K-S 검정의 p값과 카이제곱 통계량이 취하는 값의 수를 나타내었다. 그림들은 3×2 행렬 형태로 배열되었는데 첫 번째 열은 5×5 분할표에 대한 결과이며 두 번째 열은 10×10 분할표에 대한 결과이다. 또한 첫 번째 행은 기대도수가 1인 경우, 두 번째와 세 번째 행은 기대도수가 2, 4인 경우에 대한 결과이다.

<표 1>에 의하면 몇 개의 그림에서 카이제곱 통계량이 취할 수 있는 값이 많지 않아 생기는 이산성(discreteness)이 나타나고 있다. 특히 5×5 분할표의 경우 총 가능한 분할표의 수는 근사적으로 2.6×10^9개나 되지만 취하는 값이 19개이기 때문에 이산성이 아주 두드러지게 나타나고 있다. 또한 기대도수가 1인 10×10 분할표와 기대도수가 2인 5×5 분할표에도 이산성이 나타나고 있다. 이산성이 나타나는 이 그림들에서는 기대직선에서의 이탈이 나타나고 있으며 연속형 분포에 대한 검정 방법인 K-S 검정 방법에서는 이 점이 더욱 부각되어 p값이 모두 0으로 나왔다. 이산성이 두드러진 이런 경우에는 대표본 분포에 근거한 카이제곱 검정을 사용하면 보수성(conservativeness)이 나타나기 때문에 정확검정을 사용하거나 최소한 모의실험을 통해 생성된 소표본 분포에 근거한 검정을 수행해야 할 것이다. 이산성이 두드러지지 않은 다른 그림에서는 기대직선에서의 이탈이 미미하게 나타났으며 이 점은 K-S 검정의 p값이 모두 0.15보다 크게 나타나는 것으로 보다 분명하게 나타나고 있다. 요약하면 기대도수가 1인 경우에는 기대직선에서의 이탈을 볼 수 있으며 이산성이 많으면 그 정도가 심하다는 것을 알 수 있다. 그러나 기대도수가 2인 경우에는 이산성에 따라 기대직선에서의 이탈정도가 차이가 있으며, 기대도수가 4인 경우에는 이산성이 거의 나타나지 않으며 직선에서의 이탈정도가 미미하여 소표본 분포와 대표본 분포가 거의 일치한다고 할 수 있다.

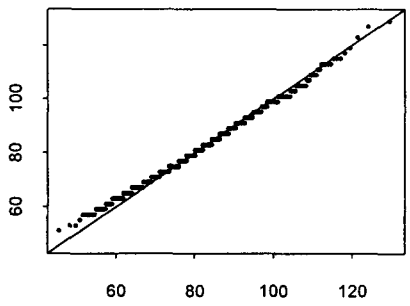
3. 결론

이 논문에서는 주변값들이 모두 주어진 이원분할표에 대한 카이제곱 검정통계량의 소표본 분포에 대해 살펴보았다. 우선 표본크기가 8 이하인 대표적인 분할표에 대해 정확분포를 계산해 보았는데 표본크기와 분할표의 크기가 조금만 커지게 되면 가능한 분할표의 개수가 기하급수적으로 늘어나기 때문에 정확분포를 계산하는 것이 상당히 어렵게 된다.

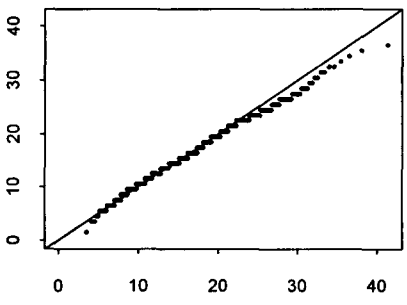
그림 1: 모의실험에 대한 카이제곱 확률도와 K-S 검정의 p값



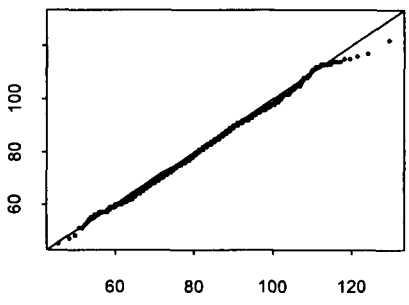
p=0, no. of chi-sq values=19



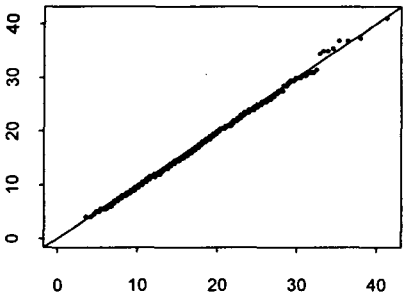
p=0, no. of chi-sq values=38



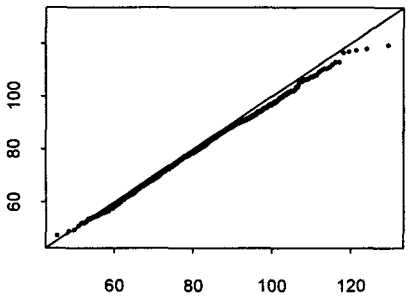
p=0, no. of chi-sq values=35



p=0.172, no. of chi-sq values=71



p=0.1604, no. of chi-sq values=61



p=0.3298, no. of chi-sq values=125

따라서 여기서는 이 어려움을 극복하기 위해 간단한 모의실험용 알로리듬을 사용하고 있다.

제시된 모의실험 알고리듬에 의해 카이제곱 통계량의 소표본 분포를 생성하였다. 실행된 모의실험은 동일한 주변값을 가지는 경우로 제한되었으며 기대도수가 1, 2, 4인 경우에 대해 실행되었다. 그 결과 기대도수가 1인 경우 카이제곱 통계량의 이산성이 나타나 기대직선에서의 이탈이 나타났으며 특히 이산성이 두드러진 5×5 분할표에서는 그 정도가 아주 심했다. 기대도수가 2인 경우에도 5×5 분할표에서는 이산성이 나타나 기대직선에서의 이탈이 나타났으나 10×10 분할표에서는 이산성이 두드러지고 않고 기대직선에서의 이탈이 심하지 않았다. 마지막으로 기대도수가 4인 경우 기대직선에서의 이탈정도가 미미하여 소표본 분포와 대표본 분포가 거의 일치함을 알 수 있었다. 따라서 동일한 주변값을 가지는 두가지 형태의 분할표에 적용된 이 모의실험에 의하면 기대도수가 4 정도만 되면 대표본 분포인 $\chi^2((r-1)(c-1)$을 사용하는데 큰 무리가 없다고 할 수 있다. 이는 기대도수가 5 이하인 것이 20% 이상이면 대표본 근사가 좋지 못하다는 통상적인 지침보다 다소 완화된 조건에서 대표본 근사가 큰 무리가 없다는 것을 시사해주는 것이다.

참고문헌

1. 허명회 (1997). 2원 분할표의 소표본 검증법. 응용통계연구 10, 339-352.

2. Alalouf, I. S. (1987). The Chi-Squared Test with Both Margins Fixed. *Communications Statistics - Theory and Methods* 16, 29-43.

3. Freeman G. H. and Halton, J. H. (1951). Note on an Exact treatment of Contingency, Goodness of Fit and Other Problems of Significance. *Biometrika* 38, 141-149.

4. Gail, M. H. and Mantel, N. (1977). Counting the Number of Contingency Tables With Fixed Margins. *Journal of American Statistical Association,* 72, 859-862.

5. Mehta, C. R. and Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ contingency tables. *Journal of American Statistical Association,* 78, 427-434.

6. Park, C. (1995). Some Remarks on the Chi-Squared Test with Both Margins Fixed. *Communications Statistics - Theory and Methods* 24, 653-661.

7. Patefield (1981). An efficient method of generating random $R \times C$ tables with given row and column totals. *Journal of Royal Statistical Society, Series, C* 30, 91-97.

8. StatXact. (1991). *StatXact: Statistical Software for Exact Nonparametric Inference,* Version 2. Cytel Software, Cambridge, Mass.

On the Small Sample Distribution and its Consistency with the Large Sample Distribution of the Chi-Squared Test Statistic for a Two-Way Contigency Table with Fixed Margins

Cheolyong Park[4]· Jaesung Choi[5]· Yonggon Kim[6]

Abstract

The chi-squared test statistic is usually employed for testing independence of two categorical variables in a two-way contingency table. It is well known that, under independence, the test statistic has an asymptotic chi-squared distribution under multinomial or product-multinomial models. For the case where both margins fixed, the sampling model of the contingency table is a multiple hypergeometric distribution and the chi-squared test statistic follows the same limiting distribution.

In this paper, we study the difference between the small sample and large sample distributions of the chi-squared test statistic for the case with fixed margins. For a few small sample cases, the exact small sample distribution of the test statistic is directly computed. For a few large sample sizes, the small sample distribution of the statistic is generated via a Monte Carlo algorithm, and then is compared with the large sample distribution via chi-squared probability plots and Kolmogorov-Smirnov tests.

Key words and Phrases: Kolmogorov-Smirnov test, Monte Carlo algorithm, multiple hypergeometric distribution, test of independence

[4]Associate Professor, Department of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Ku, Taegu 704-701, Korea

[5]Professor, Department of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Ku, Taegu 704-701, Korea

[6]Assistant Professor, Department of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Ku, Taegu 704-701, Korea