A Monte Carlo Comparison of the Small Sample Behavior of Disparity Measures

홍종선[1], 정동빈[2], 박용석[3]

Abstract

소표본 분할표 자료에서 적합도 검정통계량들의 카이제곱 근사 적용 가능에 대하여 많은 연구가 진행되었다. 소표본에서 세 가지 검정 통계량(피어슨 카이제곱 X^2, 일반화 가능도비 G^2, 그리고 역발산 $I(2/3)$ 검정통계량)에 관하여 비교한 Rudas(1986)의 연구를 확장하여, 최근에 제안된 차이측도($BWHD(1/9)$, $BWCS(1/3)$, $NED(4/3)$ 검정통계량)를 포함시켜 비교 분석하였다. 독립모형의 이차원 분할표, 조건부 독립모형과 한 변수 독립 모형을 따르는 삼차원 분할표에 대한 모의실험을 통하여 생성된 90과 95 백분위수와 이에 대응하는 95% 신뢰구간을 살펴보고 실제 백분위수와 비교하였다. 그 결과 X^2, $I(2/3)$, 그리고 $BWHD(1/9)$ 검정통계량이 유사한 결과를 나타내었고 이 통계량들이 기존에 제안된 검정통계량들보다 적은 표본크기에서도 카이제곱 근사방법에 적용 가능함을 발견하였다.

Keywords ; 가중 카이제곱 통계량, 가중 헬링거거리 통계량, 일반화 가능도비 통계량, 역발산 통계량, 음의 지수차이 통계량.

1. 서론

범주형 자료에 관하여 가설 검정하는 것은 카이제곱분포의 가장 중요한 응용방법중 하나이다. 분할표 자료를 설명하는 확률분포가 귀무가설 모형에 적합 하는 지를 검정하는 통계량으로는 피어슨 카이제곱 검정통계량과 일반화 가능도비 검정통계량이 잘 알려져 있다.

본 연구에서는 Rudas(1986)의 연구를 발전시켜 X^2, G^2, $I(2/3)$통계량, 그리고 $BWHD(1/9)$, $BWCS(1/3)$, 그리고 $NED(4/3)$통계량이 카이제곱분포에 근사함을 살펴보기 위하여 몬테칼로 방법의 반복실험을 이용하여 비교 연구하고자 한다. Rudas(1986)가 연구한 방법과 동일하게 7종류의 이차원 분할표와 6종류의 삼차원 분할표를 사용하였는데 이차원 분할표는 독립성 모형([A][B]모형)을 그리고 삼차원 분할표에서는 세번째 변수가 주어질 때 첫번째 변수와 두번째 변수가 독립인 조건부 독립모형([AC][BC]모형)을 고려하였다. 본 연구에서는 삼차원 분할표 자료에서 조건부 독립모형 외에 한 변수 독립모형([AB][C]모형)에 적합한 4가지 경우의 분할표를 추가적으로 고려하였다. 그리고 각 분할표 자료의 크기가 전체 칸의 수의 2배 또는 3배에 해당하는 소표본에 대하여, 카이제곱분포의 90과 95 백분위수와 모의실험한 90과 95 백분위수 그리고 이에 대응하는 각각의 95% 신뢰구간과의 관계를 살펴보면서 카이제곱분포의 90과 95 백분위수 수준에서 검정통계량의 적용 가능성을 살펴본다.

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수

2) (210-702) 강원도 강릉시 지변동 123, 강릉대학교 자연과학대학 정보통계학과, 부교수

3) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 대학원

2. 연구방법 및 결과

앞에서 언급한 세 가지 검정 통계량인 X^2, G^2, $I(2/3)$통계량과 차이측도 통계량 $BWHD(1/9)$, $BWCS(1/3)$, 그리고 $NED(4/3)$통계량에 대한 소표본에서의 특성은 17가지의 분할표를 이용하여 연구하였다. 이들 17가지의 분할표 중 7가지는 이차원이고 10가지는 삼차원이다. 이차원 분할표는 독립인 모형([A][B]모형)이고 삼차원 분할표 중 6가지는 세번째 변수의 주변확률이 주어졌을 때 첫번째 변수의 주변확률과 두번째 변수의 주변확률이 독립인 조건부 독립모형([AC][BC]모형)이며 나머지 4가지는 한 변수 독립모형([AB][C]모형)을 따르는 분할표 자료이다. 이차원 분할표는 2개의 일차원 주변확률표(marginal probability table)를 이용하여 생성되었다. 조건부 독립모형을 따르는 삼차원 분할표는 첫 번째와 세번째 변수 그리고 두번째와 세번째 변수의 결합 확률표를 포함하는 2개의 이차원 주변확률표를 이용하고 한 변수 독립모형을 따르는 삼차원 분할표는 첫번째와 두번째 변수의 결합확률표와 세번째 변수의 주변확률표를 이용하여 생성하였다. 모형의 자유도는 1, 2, 3, 4, 5, 10, 20으로 이차원과 삼차원 분할표 자료에 대한 자유도를 공유하게 설계하였다. 설정된 표본크기의 분할표들에 대하여 1,000개의 표본들을 생성하고 앞에서 언급한 6가지의 통계량을 1,000개의 표본들에 대하여 계산하고 값을 구한다.

본 연구에서 추가된 한 변수 독립모형에서도 Rudas(1986)의 연구와 유사하게, 소표본에서 X^2 통계량이 G^2 통계량보다 적절하며 $I(2/3)$ 통계량이 X^2 통계량과 유사함을 발견하였다. $BWHD(1/9)$ 통계량이 소표본에서 $BWCS(1/3)$와 $NED(4/3)$ 통계량보다 카이제곱 근사의 사용이 적절하다는 것을 발견하였다. 90 백분위수에 대한 결과는 95 백분위수에 대한 결과보다 세 통계량들 간의 유사성이 존재함을 살펴볼 수 있다. 그러나 이 유사성의 차이는 두 백분위수 간의 결과 차이는 크지 않다고 판단할 수 있다. $I(2/3)$ 통계량보다는 X^2 통계량이 카이제곱 근사에 적절하며 X^2 통계량보다는 $BWHD(1/9)$ 통계량이 더욱 적절함을 유도하였으며, $BWHD(1/9)$ 통계량보다는 $I(2/3)$이 그리고 $I(2/3)$ 통계량보다는 X^2 통계량이 카이제곱 근사에 적절하며, $I(2/3)$ 통계량보다 X^2과 $BWHD(1/9)$이 더욱 적절하다고 추론할 수 있다. 그러므로 모든 결과를 바탕으로 X^2, $I(2/3)$, $BWHD(1/9)$ 통계량이 나머지 통계량(G^2, $BWCS(1/3)$, $NED(4/3)$)보다 소표본에서 카이제곱 근사에 적절하다고 결론내릴 수 있다.

참고문헌

[1] Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society*, B, 46, 440-464.

[2] Jeong, D. and Sarkar, S. (2000). Negative exponential disparity family based goodness-of-fit tests for multinomial models, *Journal of Statistical Computation and Simulation*, 65, 43-61.

[3] Rudas, T. (1984). Testing goodness-of-fit of log-linear models based on small samples : a Monte Carlo study, *Colloquia Mathematica Societas J nos Bolyai*, 45, Goodness-of-fit, North-Holland.

[4] Rudas, T. (1986). A Monte Carlo comparison of the small sample behaviour of the Pearson, the Likelihood Ratio and the Cressie-Read Statistic, *Journal of the Statistical Computation and Simulation*, 24, 107-120.