응용통계연구 제10권 2호
1997년 9월 pp. 339-352

2원 분할표의 소표본 검증법

허 명 회[1)]

요 약

소표본으로부터 형성되는 2원 분할표에는 빈도가 작은 칸들이 적지 않기 때문에 대표본이론에 근거한 카이제곱 검증 등 기존 통계적 방법의 적용이 적절하지 않은 수가 있다. 이런 경우에 한 대안으로서 정확검증법(exact tests)이 개발되어 있으나 이것이 너무 많은 계산을 요구하므로 사용하기가 쉽지 않다. 본 연구는 소표본 2원 분할표에서, 단순한 몬테칼로 알고리즘에 의한 행 균일성 가설의 카이제곱 임의화 검증법(randomization test)을 제안하고 튜키(Tukey) 형의 행간 다중비교법을 제안한다. 아울러 열 범주가 순서형인 2원 분할표에 대하여도 유사한 방법론을 적용할 수 있음을 밝힌다.

1. 행간 균일성 검증

IxJ 분할표 $\{n_{ij},\ i=1,\cdots,I;\ j=1,\cdots,J\}$에서 행간 균일성은 통상적으로 피어슨의 카이제곱방법으로 검증된다. 즉 행간 균일성의 귀무가설 하에서 검증통계량

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} (n_{ij} - E_{ij})^2 / E_{ij},\quad E_{ij} = n_{i+}n_{+j}/n \tag{1}$$

이 대표본 근사론에 의해 자유도가 $(I-1)(J-1)$인 χ^2 분포를 따른다는 사실을 이용하게 된다. 그러나 기대빈도 E_{ij}가 5이하인 칸들의 수가 20%를 넘는 경우 대표본 근사는 적절하지 못한 것으로 알려져 있다 (Cochran, 1954).

소표본 자료의 경우 피셔의 정확검증(Fisher's exact test)을 사용할 수 있는데 이것은 정해진 행합계와 열합계를 갖는 모든 임의의 2x2 표 각각에 대한 조건부 확률을 이용하여 관측된 2원 분할표의 p 값을 산출하는 검증법이다. 행합계와 열합계가 주어졌을 때, 2원 분할표에 대한 조건부 확률은

$$p(n_{11}, n_{12}, n_{21}, n_{22} \mid n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{\binom{n_{+1}}{n_{11}}\binom{n_{+2}}{n_{12}}}{\binom{n}{n_{1+}}},$$

즉

$$p(n_{11}, n_{12}, n_{21}, n_{22} \mid n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{n_{+1}!\, n_{+2}!\, n_{1+}!\, n_{2+}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \tag{2}$$

---

1) 고려대학교 정경대학 통계학과 교수. [136-701] 서울특별시 성북구 안암동 5가 1.
* 1996년도 농업특정연구과제 지원을 받음.

이다. 피셔의 정확검증을 일반화하여 IxJ 표에 적용시킬 수 있는데 (예컨대 Agresti (1990, 3.5절)과 Lehmann (1975, 7.4절) 참조), 행합계와 열합계에 조건화하여 일반화 초기하확률인

$$p(\{n_{ij}\} \mid \{n_{i+}, n_{+j}\}) = \frac{\prod_{j=1}^{J} \binom{n_{+j}}{n_{1j}, \cdots, n_{Ij}}}{\binom{n}{n_{1+}, \cdots, n_{I+}}}$$

가 얻어지기 때문이다. 정리하면

$$p(n_{ij} \mid \{n_{i+}, n_{+j}\}) = \frac{(\prod_{i=1}^{I} n_{i+}!)(\prod_{j=1}^{J} n_{+j}!)}{n! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!} \tag{3}$$

이 된다. 물론 (3)의 특수한 경우로서 (2)가 얻어진다.

그런데 일반화 피셔 정확검증은, 첫째, 개별 2원 분할표들이 귀무가설로부터 이탈되는 정도를 관측확률의 크기로 함으로써 (1)을 기준으로 하는 카이제곱 검증과 일치하지 않고 (Agresti, 1992), 둘째, 분할표 및 자료의 크기가 커질수록 가능한 2원 분할표의 총 수가 매우 크므로 계산문제가 상당해지는 등의 문제를 야기시킨다. (이 방법은 윈도우 버전(Release 6.08+)의 SAS PROC FREQ에 구현되어 있는데 자료크기가 작지 않은 경우에는 엄청난 계산 시간을 소모한다. 또한, 특수목적의 소프트웨어로서 StatXact (1991)가 개발되어 있다.)

이러한 문제를 해결하고자 본 연구가 제안하고자 하는 일반 2원 분할표의 소표본 카이제곱 검증법은 임의 2원 분할표 생성을 위한 다음 몬테칼로 알고리즘에 기초한다.

- 단계 1: 2개의 원자료 열을 만든다. 즉

  행 범주값 $\rightarrow \vec{x} : n\times 1$, 열 범주값 $\rightarrow \vec{y} : n\times 1$.

- 단계 2: $\vec{y}$ 를 임의순열화하여 $\vec{y}^{*}$ 를 생성시킨다.

- 단계 3: $\vec{x}$와 $\vec{y}^{*}$ 로부터 2원 분할표를 구축한다.

이 알고리즘으로 (3)의 확률이 얻어진다는 것을 다음과 같이 논증할 수 있다.

- 단계 1에서 $\vec{x}$ 를 편의상 $(1,\cdots,1,2,\cdots,2,\cdots,I,\cdots,I)$ 라고 놓고 시작하기로 한다. 따라서 전체 가능한 경우의 수는 $\vec{y}$의 임의순열 수인 $n!$ 이다.

- y=1인 개체가 n_{+1}개 있는바 그 가운데 n_{11}개를 선발하여 x=1과 결합시키고 y=2인 개체 n_{+2}개 중에서 n_{12}개를 선발하여 x=1과 결합시킨다. 이러한 과정을 $y=J$인 개체 n_{+J}개 중에서 n_{1J}개를 선발하여 x=1과 결합시키는 것까지 계속한다. 이러한 선발 과정에서 생기는 경우의 수는

  $$\binom{n_{+1}}{n_{11}}\binom{n_{+2}}{n_{12}}\cdots\binom{n_{+J}}{n_{1J}}$$

이다. 그런데 선발된 개체들을 (그 숫자는 $n_{1+}(=n_{11}+n_{12}+\cdots+n_{1J})$) 서로 자리바꿈 시켜도 결과되는 2원 분할표가 동일하므로 결국 경우의 수는

$$n_{1+}!\binom{n_{+1}}{n_{11}}\binom{n_{+2}}{n_{12}}\cdots\binom{n_{+J}}{n_{1J}}$$

이다.

- y=1인 개체가 $n_{+1}-n_{11}$개 남았는바 그 가운데 n_{21}개를 선발하여 x=2와 결합시키고 y=2인 개체 중 남은 $n_{+2}-n_{12}$개 중에서 n_{22}개를 선발하여 x=2와 결합시킨다. 이러한 과정을 $y=J$인 개체 $n_{+J}-n_{1J}$개 중에서 n_{2J}개를 선발하여 x=2와 결합시키는 것까지 계속한다. 이 과정에서 생기는 경우의 수는

$$\binom{n_{+1}-n_{11}}{n_{21}}\binom{n_{+2}-n_{12}}{n_{22}}\cdots\binom{n_{+J}-n_{1J}}{n_{2J}}$$

이다. 그런데 선발된 개체들을 (그 숫자는 $n_{2+}(=n_{21}+n_{22}+\cdots+n_{2J})$) 서로 자리바꿈 시켜도 결과되는 2원 분할표가 동일하므로 결국 경우의 수는

$$n_{2+}!\binom{n_{+1}-n_{11}}{n_{21}}\binom{n_{+2}-n_{12}}{n_{22}}\cdots\binom{n_{+J}-n_{1J}}{n_{2J}}$$

이다.

- 이러한 과정을 $x=I$에 대하여까지 계속 진행하면, 2원 분할표 $\{n_{ij}\}$가 결과되는 총 순열의 수는

$$\prod_{i=1}^{I}\left\{n_{i+}!\prod_{j=1}^{J}\binom{n_{+j}-n_{1j}-\cdots-n_{i-1,j}}{n_{ij}}\right\}=\frac{(\prod_{i=1}^{I}n_{i+}!)(\prod_{j=1}^{J}n_{+j}!)}{}$$

가 된다.

- 특정 2원 분할표에 대한 조건부 확률은 이것을 총 순열의 수 $n!$로 나눈 것이다.

증명 끝.

따라서 소표본 카이제곱검증에서의 정확한 p 값은 다음과 같이 추정가능하다 (임의화 검증(randomization test)의 방식에 대한 일반적인 설명에 대하여는 Good(1994)를 참조하라).

- 단계 1: $\vec{x}$와 임의순열 $\vec{y}^*$로부터 2원 분할표를 구축하여 카이제곱통계량 X^{*2}을 계산한다.
- 단계 2: 단계 1을 N번 반복하면서 $X^{*2}\geq X^2$인 상대적 빈도 p를 산출한다.
  여기서 X^2은 원자료에서 얻어진 카이제곱 통계값이다.

단계 2에서의 N은 상대빈도가 상당한 정밀도를 얻을 수 있을 만큼 큰 수이다. 예컨대 N = 10000으로 잡으면 p 값의 추정치 p에 대한 최대 표준오차 및 추정 표준오차(s.e.)는 각각

$$0.5/\sqrt{N}=0.0050 \text{ 과 } \sqrt{p(1-p)/N}$$

이다. 그리고, 2x2 분할표에 대하여는 단계 1에서 검증통계량으로 식 (1)의 카이제곱을

$$\chi_{11} = (n_{11} - E_{11})/\sqrt{E_{11}}$$

로 대치함으로써 단측검증도 가능하다.

이하 사례적용에서는 몬테칼로 시행 수 N을 10000으로 하였다. 왜냐하면 그 경우 95% 신뢰수준하에서의 최대 표준오차를 0.01, 즉 1%로 할 수 있는데 이 정도면 경험적으로 p 값에 대하여 충분히 정확한 신뢰한계를 얻을 수 있기 때문이다 (물론, 아래의 사례 3과 같은 예외가 있다). 요구되는 계산은 모두 SAS/IML로 처리하였다.

사례 1) Good, P. (1994) *Permutation Tests*, 85-86.

$$\text{분할표} = \begin{pmatrix} 0 & 1 & 0 \\ 8 & 1 & 8 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \qquad X^2 = 22.10$$

대표본 근사론 : 자유도 16의 카이제곱 분포, 근사 p 값 0.1400
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0283 (s.e. = 0.0017)
참고 : StatXact 제공 p 값 0.0269

사례 2) Agresti, A. *et al.* (1979) *Psychometrika* 44, 75-83.

$$\text{분할표} = \begin{pmatrix} 10 & 1 & 6 \\ 3 & 5 & 0 \\ 5 & 0 & 1 \end{pmatrix} \qquad X^2 = 14.81$$

대표본 근사론 : 자유도 4의 카이제곱 분포, 근사 p 값 0.0051
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0034 (s.e. = 0.0006)
참고 : StatXact 제공 p 값 0.0038

사례 3) Agresti, A. (1990) *Categorical Data Analysis*. 64.

$$\text{분할표} = \begin{pmatrix} 25 & 25 & 12 \\ 0 & 1 & 3 \end{pmatrix} \qquad X^2 = 6.96$$

대표본 근사론 : 자유도 2의 카이제곱 분포, 근사 p 값 0.0309
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0490 (s.e. = 0.0022)
참고 : StatXact 제공 p 값 0.0516

사례 4) 千國외 5인 (1991)「일본축산학회보」 62(7). 660-666.

$$\text{분할표} = \begin{pmatrix} 6 & 10 & 9 \\ 13 & 5 & 2 \\ 7 & 2 & 1 \\ 2 & 3 & 1 \end{pmatrix} \qquad X^2 = 11.87$$

대표본 근사론 : 자유도 6의 카이제곱 분포, 근사 p 값 0.0648
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0597 (s.e. = 0.0024)
참고 : StatXact 제공 p 값 0.0613

사례 3에서와 같이 p 값에 관한 95% 수준의 신뢰구간이 (0.0446, 0.0534)로서 0.05에 걸쳐있는 경우, 논란의 여지가 있기는 하나 필자는 "유의수준 α= 5%에서 귀무가설을 기각하지 못한다"는 보수적 입장에 서고 싶다. 왜냐하면 제시된 귀무가설에 배반되는 경험적 증거가 충분하지 않는 한 귀무가설을 버릴 수 없다는 것이 유의성 검증의 기본 생각이기 때문이다.

2. 행간 다중비교법

분산분석에서와 유사하게 2원 분할표 자료의 분석에서도 다중비교가 필요하다. 먼저, 2xJ 분할표에서

행 1 : $(n_{11}, n_{12}, \cdots, n_{1J})/n_{1+}$, 행 2 : $(n_{21}, n_{22}, \cdots, n_{2J})/n_{2+}$

사이에 유의한 차이가 있는지를 검증하여 보자. 행간 피어슨의 카이제곱 거리

$$d_\chi^2(1,2) = \sum_{j=1}^{J} \left(\frac{n_{1j}}{n_{1+}} - \frac{n_{2j}}{n_{2+}}\right)^2 / \frac{n_{+j}}{n}$$

와 카이제곱통계량 (1) 사이에는

$$X^2 = n_{1+}n_{2+}/n \cdot d_\chi^2(1,2) \tag{4}$$

라는 관계가 성립한다 (증명은 부록 1 참조). 따라서 카이제곱거리를 약간만 보정하면 행간 차이 검증이 가능함을 알 수 있다. 물론 행의 수 I가 2인 경우에 있어서는 행간 다중비교가 필요하지 않다.

IxJ 분할표에 있어서는 행 r과 행 s 사이의 다항비율

행 r : $(n_{r1}, n_{r2}, \cdots, n_{rJ})/n_{r+}$, 행 s : $(n_{s1}, n_{s2}, \cdots, n_{sJ})/n_{s+}$

사이의 피어슨 카이제곱 거리

$$d_\chi^2(r,s) = \sum_{j=1}^{J} \left(\frac{n_{rj}}{n_{r+}} - \frac{n_{sj}}{n_{s+}}\right)^2 / \frac{n_{+j}}{n}$$

와 카이제곱통계량 (1) 사이에 다음의 관계가 있음을 어렵지 않게 보일 수 있다 (증명은 부록 1 참조).

$$X^2 = \sum_{r=1}^{I} \sum_{s=1}^{I} (n_{r+} + n_{s+})/2n \cdot n_{r+}n_{s+}/(n_{r+} + n_{s+}) \cdot d_\chi^2(r,s). \tag{5}$$

따라서 개별적으로 자유도 J-1의 카이제곱분포를 근사적으로 따르는

$$C(r,s) \equiv n_{r+}n_{s+}/(n_{r+} + n_{s+}) \cdot d_\chi^2(r,s), \quad r=1,\cdots,I;\ s=1,\cdots,J;\ r \neq s$$

의 가중평균으로 X^2 통계량이 표현된다. 그러므로 이것들을 다중적으로 사용하여야 하는 행

간의 다중비교를 위해서는, 분산분석에서 튜키(Tukey)의 표준화 범위 검증(studentized range test)이 그러하듯이, 최대 카이제곱 형의 통계량

$$C_M \equiv \max_{r,s} C(r,s) \tag{6}$$

를 검증통계량으로 사용해야 할 것이다. 문제는 C_M의 귀무분포(null distribution)를 구하는 것인데 통계적으로 상호간 관련되어 있는 다수 임의량들의 최대값과 관련 있으므로 해석적 접근이 불가능해 보인다. 그러므로 몬테칼로 계산에 의한 귀무분포의 추정방법을 생각해 보자.

- 단계 1: $\vec{x}$와 임의순열 $\vec{y}^*$로부터 2원 분할표를 만들어 식 (6)에 의한 C_M^*을 계산한다.
- 단계 2: 단계 1을 N번 반복하여 C_M 통계량 (6)의 귀무분포 G를 구축한다.
- 단계 3: 분포 G의 상위 100 α% 분위수 g_α를 산출하여 다중비교에 활용한다.
  즉, $C(r,s) \geq g_\alpha$이면 행 r과 s가 유의하게 다르다고 선언한다. 그렇지 않으면 행 r과 s가 유의하게 다르다고 보지 않는다.

몬테칼로 변동이 g_α의 추산에 주는 영향을 알기 위하여는 이상의 과정을 L번 독립적으로 반복하여 $\{g_{\alpha[l]},\ l=1,\cdots,L\}$ 을 얻은 뒤, 이것들의 평균 $\bar{g}_\alpha$와 이에 붙는 표준오차 $s.e.$를 구하여야 한다. 즉,

$$\bar{g}_\alpha = L^{-1}\sum_{l=1}^{L} g_{\alpha[l]}, \quad s.e. = \sqrt{(L(L-1))^{-1}\sum_{l=1}^{L}(g_{\alpha[l]}-\bar{g}_\alpha)^2}.$$

몇가지 사례에 적용해보기로 하겠다 (SAS/IML로 프로그램되었음). 다음에서 본페로니(Bonferroni) 분위수는 자유도가 J-1인 카이제곱 분포에서 상위 α/K 분위수를 말한다. 여기서 K는 모든 행의 짝 비교 수인 $I(I-1)/2$ 이다.

사례 1) 이상진 외 6인 (1994) 「한국가축번식학회지」 18, 217-228.

분할표 = $\begin{pmatrix} 60 & 0 \\ 60 & 0 \\ 55 & 5 \\ 49 & 11 \\ 30 & 30 \end{pmatrix}$ $\quad X^2 = 79.96$

대표본 근사론 : 자유도 4의 카이제곱 분포, 근사 p 값 0.000
몬테칼로 1000x10번 시행(N=1000, L=10)의 결과 : 추정 p 값 0.0000
행간 다중비교 결과 (α =0.05) : 행 1 2 3 4 5
---

* C_M 통계량의 상위 5% 분위수는 7.767로 추산됨 (s.e.=0.000).
참고: 카이제곱분포(자유도 1)의 상위 5% 분위수 3.84.,
상위 0.5% (본페로니) 분위수 7.88.

사례 2) 김상근 외 2인 (1995) 「한국가축번식학회지」 19, 129-134.

분할표 = $\begin{pmatrix} 1 & 1 & 2 & 2 & 6 & 4 \\ 4 & 5 & 4 & 4 & 3 & 3 \\ 4 & 4 & 5 & 4 & 3 & 2 \end{pmatrix}$ $\quad X^2 = 8.43$

대표본 근사론 : 자유도 10의 카이제곱 분포, 근사 p 값 0.586
몬테칼로 1000x10번 시행 결과 : 추정 p 값 0.6138 (s.e.=0.0049).
행간 다중비교 결과 (a =0.05) : 행 1 2 3 4 5
---

* C_M 통계량의 상위 5% 분위수는 13.278 (s.e.=0.109).
참고: 카이제곱분포(자유도 5)의 상위 5% 분위수 11.07.
상위 1.67% (본페로니) 분위수 13.89.

사례 3) Good, P. (1994) *Permutation Tests*, 85-86. 계속
몬테칼로 1000x10번 시행 결과 : 추정 p 값 0.0283 (s.e. = 0.0017).
행간 다중비교 결과 (a =0.05) : 행 1 2 3 4 5 6 7 8 9
---

* C_M 통계량의 상위 5% 분위수는 6.475 (s.e.=0.030).
참고: 카이제곱분포(자유도 2)의 상위 5% 분위수 5.99.
상위 0.14% (본페로니) 분위수 13.89.

사례 4) 千國외 5인 (1991) 「일본축산학회보」 62(7), .660-666. 계속
몬테칼로 1000x10번 시행 결과 : 추정 p 값 0.0597 (s.e. = 0.0024).
행간 다중비교 결과 (a =0.10) : 행 1 3 4 2
---

* C_M 통계량의 상위 10% 분위수는 7.434 (s.e.=0.080).
참고: 카이제곱분포(자유도 2)의 상위 10% 분위수 4.61,
상위 1.67% (본페로니) 분위수 8.18.

앞의 모든 사례에서 보듯이, 당연하게도, C_M 통계량의 상위 a 분위수는 카이제곱 분포(자유도 J-1)의 상위 a 분위수보다 크고 본페로니 분위수보다는 작은 값을 취한다. 한편 앞의 사례 3에서는, 카이제곱검증이 유의한 결과를 보였지만 다중비교법이 행간 차이를 구체적으로 보이지 못함으로써, 분할표에서의 행간 다중비교법이 카이제곱검증에 비하여 검증력(test power)이 다소 떨어질 것이라는 예상을 가능하게 한다. .

3. 열 범주가 순서형인 경우

열이 범주형인 경우에는 각 열 범주에 어떤 수량치를 부여한 후 "행간 균일성"의 문제를 다루는 것이 일반적이다. 수량화(점수화)의 방법에는 여러 가지가 가능하지만 본 연구에서는 열 범주를 등순위(tied rank)로 수량화하기로 하겠다 (Klotz and Teng (1977), Agresti (1992), Agresti (1996, 2.5절)). 예컨대, 2원 분할표에서 열 범주 1, 2, 3, 4의 주변빈도가 4, 3, 2, 1인 경우

열자료값 : 1, 1, 1, 1, 2, 2, 2, 3, 3, 4
순 위 : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
등 순 위 : 2.5, 2.5, 2.5 2.5, 6.0, 6.0, 6.0, 8.5, 8.5, 10

이므로, 열 범주 1이 2.5로, 범주 2가 6.0, 범주 3이 8.5, 범주 4가 10으로 수량화된다.

일반적으로, IxJ 분할표 $\{n_{ij},\ i=1,\cdots,I;\ j=1,\cdots,J\}$에 대하여는 열의 범주 j가 다음과 같이 수량화된다.

$$c_j = (\sum_{k=1}^{j-1} n_{+k} + 1 + \sum_{k=1}^{j} n_{+k})/2.$$

따라서 $c_1 \le \cdots \le c_J$가 되며, 크러스칼-왈리스(Kruskal-Wallis)형의 한 버전으로

$$F = \frac{\sum_{i=1}^{I} n_{i+}(\bar{x}_i - \bar{x})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}(c_j - \bar{x}_i)^2/(n-I)},$$

를 생각하여 보자. 여기서

$$\bar{x}_i = \sum_{j=1}^{J} n_{ij} c_j / n_{i+}\ \ (i=1,\cdots,I),\quad \bar{x} = \sum_{i=1}^{I} n_{i+}\bar{x}_i/n = (n+1)/2.$$

F의 귀무분포로서 대표본근사론에 의한 F(I-1, n-I) 분포 대신, 1절에서와 같은 몬테칼로 방법을 생각해 볼 수 있다. 다시 말하자면, 행합계와 열합계에 조건화하여 생성시킨 임의 2원 분할표들로부터 산출된 F 분포로부터 F에 대한 p 값을 추산할 수 있다는 것이다.

다음은 그런 생각을 몇몇 사례에 적용해본 것이다 (SAS/IML로 프로그램되었음).

사례 1) 김상근 외 2인 (1995) 「한국가축번식학회지」 19. 129-134. 계속
F = 3.45 , d.f. = (2, 58) , 대표본 근사 p 값 = 0.0385.
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0403 (s.e. = 0.0020).

사례 2) 千國외 5인 (1991) 「일본축산학회보」 62(7), 660-666. 계속
F = 4.17 , d.f. = (3, 57) , 대표본 근사 p 값 = 0.0097.
몬테칼로 10000번 시행 결과 : 추정 p 값 0.0089 (s.e. = 0.0009).

행간 다중비교는 일단 열이 수량화되어 있으므로 튜키(Tukey)의 다중비교법을 적용할 수 있을 것이다. 즉, 검증통계량을

$$Q = \max_{r=1,\cdots,I;\, s=1,\cdots,I} \frac{|\bar{x}_r - \bar{x}_s|}{\sqrt{\hat{\sigma}^2\, 1/2(1/n_{r+} + 1/n_{s+})}}, \qquad (7)$$

여기서 $\hat{\sigma}^2$은 오차분산추정치

로 놓자. 임계값의 산출방법으로, 대표본 근사론을 따르면 표준화범위표에서 읽기만하면 되지만, 소표본 임계값을 얻기 위해서는 행합계와 열합계에 조건화하여 생성시킨 임의 2원 분할표를 활용하는 몬테칼로 방법을 생각할 수 있다. 다음은 그런 생각을 몇몇 사례에 적용해본 것이다 (SAS/IML로 프로그램되었음).

사례 1) 김상근 외 2인 (1995) 「한국가축번식학회지」 19, 129-134. 계속
대표본근사 Q 분포 (처리수 3,자유도 60) : 5% 임계값 3.40, 10% 임계값 2.96
몬테칼로 10000번 시행 결과 : 5% 임계값 3.45, 10% 임계값 2.94
다중비교결과 (유의수준 10%) : 행 1 2 3

사례 2) 千國외 5인 (1991) 「일본축산학회보」 62(7), 660-666. 계속

대표본 근사 Q 분포 (처리수 4,자유도 60) : 5% 임계값 3.74

몬테칼로 10000번 시행 결과 : 5% 임계값 3.73

다중비교결과 : 행 1 4 2 3

열 범주뿐만 아니라 행 범주도 순서형인 경우, "행간 비균일성"은 행과 열의 연관성으로 이해된다. 여러 연관성 측도(measure of association) 중에서 이런 경우 적용될 수 있는 것은 감마(gamma, γ)이다. 그러나 γ에서는 행과 열에서 범주간 간격이 고려되지 않는다. 따라서 행과 열을 등순위(tied rank)로 수량화한 뒤 상관계수로서 연관성 측도와 검증통계량을 얻는 방법을 생각할 수 있다. 즉 행 범주 i와 열 범주 j를

$$r_i = (\sum_{k=1}^{i-1} n_{k+} + 1 + \sum_{k=1}^{i} n_{k+})/2, \quad i=1,\cdots,I$$

$$c_j = (\sum_{k=1}^{j-1} n_{+k} + 1 + \sum_{k=1}^{j} n_{+k})/2, \quad j=1,\cdots,J$$

로 수량화한 뒤, 상관계수

$$R = \frac{\sum_i \sum_j n_{ij} (r_i - \bar{r})(c_j - \bar{c})}{\sqrt{\sum_i n_{i+} (r_i - \bar{r})^2} \cdot \sqrt{\sum_j n_{+j} (c_j - \bar{c})^2}}$$

을 검증통계량으로 정한다. R의 귀무분포는 대표본근사론을 따라 변환

$$t = \sqrt{n-2} \cdot R / \sqrt{1-R^2}$$

을 취한 후 t 분포(자유도= n-2)로부터 얻어질 수도 있지만, 대안으로서 행합계와 열합계에 조건화하여 생성시킨 다수의 임의 2원 분할표로부터 관측 상관계수 R에 대한 p 값을 추산하는 몬테칼로 방법을 생각할 수 있다. 다음은 적용 예이다 (SAS/IML로 프로그램되었음).

사례 1) Agresti, A. (1990) *Categorical Data Analysis*. 64. 계속

$R=0.29$

t 분포(자유도=64) : 근사 p 값(단측) 0.0089,

몬테칼로 10000번 시행 결과 : 추정 p 값(단측) 0.0197 (s.e.=0.0014).

사례 2) 千國외 5인 (1991) 「일본축산학회보」 62(7), 660-666. 계속

$R=-0.31$

t 분포(자유도=59) : 근사 p 값(단측) 0.0072,

몬테칼로 10000번 시행 결과 : 추정 p 값(단측) 0.0075 (s.e.=0.0009).

행 범주가 순서형인 경우에 있어서 행간 다중비교는 대안가설이 행의 수준에 따라 평균적 열 수량값이 증가 또는 감소한다는 것이 된다. 따라서 앞의 (7) 대신

i) 증가의 경우, $Q_+ = \max_{1 \le r < s \le I} \frac{\bar{x}_s - \bar{x}_r}{\sqrt{\hat{\sigma}^2 1/2(1/n_{r+} + 1/n_{s+})}}$

ii) 감소의 경우, $Q_- = \max_{1\le r<s\le I} \dfrac{\bar{x}_r - \bar{x}_s}{\sqrt{\hat{\sigma}^2 1/2(1/n_{r+}+1/n_{s+})}}$

에 의한 튜키형의 다중비교법으로 해결될 수 있다.

이외에 행과 열이 모두 순서형인 경우 생각될 수 있는 검증법으로 Jonckheere- Terpstra (JT) 검증이 있겠고 (Hollander and Wolfe, 1973; Chapter 6), JT를 이용하여 앞에서와 유사하게 행간 다중비교법을 구축할 수 있을 것이다.

4. 나오며

분할표의 정확추론(exact inference)에 관하여는 Agresti (1992)의 개괄논문이 있고 Agresti의 책 (1990, Chapter 3)에도 일부 다루어져 있다. 소표본 정확추론에서 가장 문제가 되는 것은 계산상의 난점인데, 예컨대 I=J=4 (5)이고 n=20인 경우 주어진 행합계와 열합계를 갖는 2원 분할표가 최대 4만 (200만) 개가 있고 n=100이면 가능한 2원 분할표의 수는 72억 (210조) 개에 달한다 (Agresti, 1992). 따라서 단순히 모든 2원 분할표를 생성시켜 보는 것은 절대로 무리이며, 이를 해결하기 위한 방법으로 네트워크 알고리즘(Mehta and Patel, 1983)과 중요도 표집 (Mehta *et al.*, 1988) 또는 몬테칼로 알고리즘 (Agresti *et al.*, 1979) 등이 고안되었다. 전자의 경우는 정확한 p 값을 제공하기는 하나 알고리즘이 상당히 복잡하여 현실적으로는 특수 소프트웨어인 StatXact (1991)에 의존하지 않을 수 없고 후자의 경우 비균일확률인 (3)에 의존하여 임의의 2원 분할표가 생성되므로 프로그램시 상당한 복잡성이 야기된다. 반면, 본연구에서의 몬테칼로 알고리즘은 원자료를 균일확률로 재배열시켜 임의의 2원 분할표를 얻기 때문에 초급통계학 강의에서 다루어질 수 있을 정도로 단순하다 (Diaconis and Efron (1985)은 2원 분할표에 대한 또 다른 의미의 균일분포에 대하여 연구하였다).

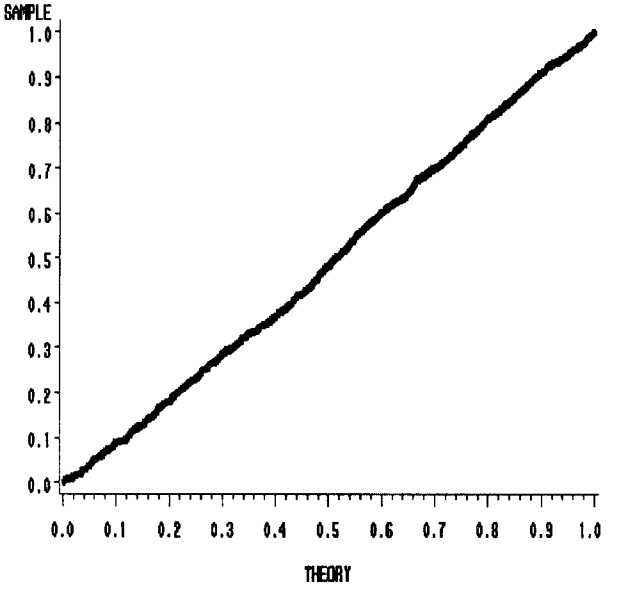
본 연구에서 제안한 고정 행합계와 열합계를 갖는 임의 2원 분할표의 생성방법은 쉽게 3원 분할표 또는 그 이상의 다차원 분할표에도 확대 적용될 수 있다. 3원 분할표에서 제1, 제2, 제3 분류변수를 X, Y, Z라고 할 때

- 단계 1: 주어진 3원 분할표로부터 3개의 원자료 열 $\vec{x}$, $\vec{y}$, $\vec{z}$: $n\times 1$을 만든다.
- 단계 2: $\vec{y}$를 임의순열화하여 $\vec{y}^*$를 생성시킨다. 독립적으로 $\vec{z}$를 임의순열화하여 $\vec{z}^*$를 생성시킨다.
- 단계 3: $\vec{x}$와 재생성된 자료열 $\vec{y}^*$, $\vec{z}^*$로부터 3원표를 구성한다

와 같이 함으로써, 각 분류변수의 주변빈도가 고정된 3원표를 임의로 생성시킬 수 있기 때문이다. 이를 이용하면 3원 분할표에 대한 로그-선형 분석에서의 소표본 임의화 검증이 가능할 것이다.

본 연구에서의 몬테칼로에 의한 2원 분할표의 소표본 카이제곱 검증은 크게 보면 순열 검증 (permutation test) 또는 임의화 검증(randomization test)이라고 할 수 있다. 이에 관하여는 Fisher (1935, 3장)가 효시이며 Cox and Hinkley (1974, 6장)와 Good (1994) 등이 일부 이론과 방법을 설명하였다. 그러나 본 연구에서 제안한 2원 분할표의 "소표본 몬테칼로" 카이제곱이

점근적으로 전통적인 카이제곱 분포를 따르는지는 아직 논증하지 못하였다. 단지 1절의 사례 2, 3, 4와 연구자가 시도해본 여러 사례로부터 그러할 가능성이 다분하다고 유추할 뿐이다. 예컨대 1절 사례 4의 4x3 표로부터 생성되는 임의 카이제곱값 999개를 자유도 6의 카이제곱분포에 견주어 p-p 플롯을 그려보았더니 <그림 1>과 같이 나타났다.



<그림 1> p-p 플롯 : 자유도 6의 카이제곱과의 비교 (세로: 표본, 가로: 이론)

참고문헌

[1] Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York..

[2] Agresti, A. (1992). "A survey of exact inference for contingency tables," *Statistical Science* 7, 131-177.

[3] Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* Wiley, New York..

[4] Agresti, A., Wackerly, D., and Boyett, J. (1979). "Exact conditional tests for cross-classifications: Approximations of attained significance levels," *Psychometrika* 44, 75-83.

[5] Cochran, W.G. (1954). "Some methods of strengthening the common X^2 tests," *Biometrics* 10, 417-451.

[6] Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

[7] Diaconis, P., and Efron, B. (1985). "Testing for independence in a two-way table: New interpretations of the chi-square statistic (with discussions)," *The Annals of Statistics* 13, 845-913.

[8] Fisher, R.A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

[9] Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, New York.

[10] Greenacre, M., and Hastie, T. (1987). "The geometric interpretation of correspondence analysis," *Journal of American Statistical Association* 82, 437-447.

[11] Hollander, M., and Wolfe, D.A. (1973). *Nonparametric Statistical Methods.* Wiley, New York.

[12] Klotz, J., and Teng, J. (1977). "One-way layout for counts and the exact enumeration of the Kruskal-Wallis H distribution with ties," *Journal of American Statistical Association* 61, 772-787.

[13] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Hoden-Day, San Francisco.

[14] Mehta, C.R., and Patel, N.R. (1983). "A network algorithm for performing Fisher's exact test in r x c contingency tables," *Journal of American Statistical Association* 78, 427-434.

[15] Mehta, C.R., Patel, N.R., Senchaudhuri, P. (1988). "Importance sampling for estimating exact probabilities in permutational inference," *Journal of American Statistical Association* 83, 999-1013.

[16] StatXact (1991). *StatXact: Statistical Software for Exact Nonparametric Inference,* Version 2. Cytel Software, MA: Cambridge.

[17] 김상근 외 2인 (1995). "돼지 분할수정란 및 미성숙란의 생존율에 관한 연구", 「한국가축번식학회지」 19(2), 129-134.

[18] 이상진 외 6인 (1994). "핵치환에 의한 Clone Animal의 생산에 관한 연구", 「한국가축번식학회지」 18(3), 217-228.

[19] 千國幸一 외 5인 (1991). "PCR法を用いた牛成長ホルモン遺傳子127番アミノ酸 部位 鹽基配列の多型檢出", 「日本畜産學會報」 62(7), 660-666.

부록 1: 카이제곱통계량 X^2과 행간 거리 사이의 관계 (4)와 (5)의 증명

식 (4)는 (5)의 특수한 경우이므로 (I=2), 여기서는 식 (5) 만을 보이도록 한다.
행 i의 프로파일(row profile)을 $1\times J$ 벡터 a_i^t로 나타내자. 즉,

$$a_i^t = (n_{i1}, n_{i2}, \cdots, n_{iJ})/n_{i+}, \quad i=1,\cdots,I$$

이다. 그리고 행 프로파일들 각각에 가중치 $w_1,\cdots,w_I$ ($w_i = n_{i+}/n$, $i=1,\cdots,I$)를 두어 가중평균낸 평균 행 프로파일(average row profile)을 c^t라고 하자. 즉,

$$c^t = w_1 a_1^t + \cdots + w_I a_I^t \ , \ i=1,\cdots,I; \quad \sum_i^I w_i = 1$$

이다. 그러면 행 r과 행 s 사이의 피어슨 카이제곱 거리 $d_\chi^2(r,s)$가

$$d_\chi^2(r,s) = (a_r - a_s)^t D_c^{-1}(a_r - a_s), \ r=1,\cdots,I; \ s=1,\cdots,I$$

로 표현된다. 또한 카이제곱통계량 X^2이

$$X^2 = n\sum_{i=1}^I w_i(a_i - c)^t D_c^{-1}(a_i - c)$$

와도 일치하므로 (Greenacre and Hastie, 1987),

$$X^2 = n\{\sum_{i=1}^I w_i a_i^t D_c^{-1} a_i - c^t D_c^{-1} c\} = n\{\sum_{i=1}^I w_i a_i^t D_c^{-1} a_i - 1\}$$

이다 ($\because D_c^{-1}c = e$, $c^t e = 1$. 여기서 $e^t = (1,\cdots,1)$). 한편,

$$\begin{aligned}
\sum_{r=1}^I \sum_{s=1}^I w_r w_s d_\chi^2(r,s) &= \sum_{r=1}^I \sum_{s=1}^I w_r w_s (a_r - a_s)^t D_c^{-1}(a_r - a_s) \\
&= \sum_{r=1}^I \sum_{s=1}^I \{2 w_r w_s a_r^t D_c^{-1} a_r - 2 w_r w_s a_r^t D_c^{-1} a_s\} \\
&= \sum_{r=1}^I \{2 w_r a_r^t D_c^{-1} a_r - 2 w_r a_r^t D_c^{-1} c\} \\
&= 2\sum_{r=1}^I \{w_r a_r^t D_c^{-1} a_r - 1\}
\end{aligned}$$

이므로

$$X^2 = n/2 \cdot \sum_{r=1}^I \sum_{s=1}^I w_r w_s d_\chi^2(r,s)$$

이 된다. 이것이 식 (5)와 일치하는 표현임을 쉽게 확인할 수 있다.

Small Sample Tests for Two-way Contingency Tables

Myung-Hoe Huh [2)]

Abstract

Chi-square test based on large sample theory is inappropriate for testing the row homogeneity in two-way contingency table with several sparse cells. For that case, exact testing methods has been developed in the literature and implemented in StatXact (1991). However, considerable computing time is inevitable for moderate size tables. So, Monte Carlo approximation is recommended frequently.

In this study, we propose a simple algorithm for generating two-way random tables with fixed row and column margins for small sample chi-square test. Also, we develop "Tukey-type" method for multiple between-rows comparisons. The proposed small sample method is computationally intensive since its null distribution of the test statistic - the maximum of chi-squares - is to be obtained from a Monte Carlo procedure, case by case.

For two-way tables with ordered column categories, similar testing methods for the homogeneity of all rows and for the multiple comparison of rows are proposed, with Pearson's chi-square replaced by Kruskal-Wallis statistic.

2) Professor, Dept. of Statistics, Korea University. Anam-dong 5-1, Seoul 136-701, Korea.