응용통계연구 제 17권 1호, 2004년, pp. 89-103

모집단 부분정보가 주어진 상황에서의 분할표 독립성 검정*

이광진[1)]

요 약

2차원 분할표 형태의 자료에 대해 두 범주형 변수들간의 독립성을 검정함에 있어, 만일 두 변수 각각의 모분포가 이미 완전히 알려진 경우라면 이 알려진 정보를 충족할 수 있도록 분할표 자료를 보정한 후 보정된 분할표 자료에 대해 전통적인 카이제곱 검정법을 적용하는 것이 더 타당함을 논증한다. 그리고 이에 근거한 제약상황 카이제곱 독립성 검정법을 유도하고 모의실험을 통해 전통적인 무보정 카이제곱 검정법과 비교한다.

주요용어: 분할표, 모집단 부분정보, 독립성 검정법, 자료보정, 사후 가중법

1. 서론

「인구주택총조사」와 같은 관청 주도의 전수조사들이 그 종류가 다양해지고 또 주기적으로 이루어지면서 지역, 성, 학력, 연령 등과 같은 인구사회적 변수들에 대해서는 모집단 정보들이 이미 상당히 많이 알려져 있다. 또한 기업, 학교 및 각종 단체 등에서 고객 또는 구성원들에 관한 다양한 종류의 항목들이 체계적으로 수집·관리되면서 이제 우리는 매우 다양한 집단에 대해 상당히 많은 변수들에 관한 정확한 정보들을 쉽게 활용할 수 있게 되었다. 이러함에도 불구하고 이미 알려진 모집단 정보들이 실제 통계분석의 과정에서는 충분히 활용되고 있지 못함은 '정보의 최대한 활용'을 추구하는 통계학적인 입장에서는 매우 안타까운 현실이라고 할 수 있다.

한편, 어떤 표본추출 방법을 사용하든 모집단에 관한 주어진 정보와 어느 정도의 차이를 보이는 표본이 얻어질 가능성은 상존하며, 실제로 상당한 차이를 보이는 경우도 가끔 발생되고 있다. 그럼에도 불구하고 많은 연구자들이 입수된 자료에 내재된 이러한 자료의 편향성을 단지 랜덤 메카니즘에 의해 얼마든지 나타날 수 있다는 식으로 무시한 채 아무런 보정의 절차도 없이 주어진 표본에만 의존한 통계적 추론과정을 통해 결론을 도출한 경우도 흔히 찾아볼 수 있다.

물론 모평균이나 모비율과 같은 일부 모수의 추정문제에서는 사후적으로 개별자료에 가중치를 주는 방법으로 표본과 모집단의 차이를 없애는 사후보정법이 그나마 어느 정도 적용되고 있기는 하다. 하지만 검정의 문제에까지 사후보정법이 제대로 사용된 경우는 거의 찾아보기 힘들다고 할 수 있다. 이는 아마도 자료보정을 고려한 검정방법들이 충분히 제시되지 못했거나, 아니면 기존의 잘 확립된 검정방법을 보정된 자료에 적용할 경우 검정통계량의 분포가 어떻게 변하는지에 관한 연구결과가 미흡했기 때문일 것으로 여겨진다.

* 이 논문은 1998년 목원대학교 교내연구비에 의하여 연구되었음.

1) (302-729) 대전시 서구 도안동 800, 목원대학교 사회과학대학 정보통계학과, 부교수
E-mail: leekj@mokwon.ac.kr

이에 본 연구에서는 행과 열 두 주변분포가 완전히 알려진 모집단에서 얻어진 2차원 분할표 자료에 대한 독립성의 검정에서, 우선 자료에 내재된 편향성을 주어진 정보에 근거하여 사후보정한 후 이 보정된 자료에 대해 적용할 수 있는 독립성 검정방법의 개발이 필요하다는 인식하에 이를 시도하며, 이 결과에 바탕을 둔 제약상황 독립성 검정법을 제시하고 그 타당성을 보이고자 한다.

2. 문제제기

A항아리 속에 아주 많은 개수의 반쪽 공들이 들어 있는데 그 중 $100\alpha\%$ 만큼에는 숫자 '1'이, 나머지에는 숫자 '2'가 적혀 있다고 하자. 그리고 B항아리 속에도 A항아리 속에 있는 개수만큼의 반쪽 공들이 들어 있는데 그 중 $100\beta\%$ 만큼에는 '흰색'이, 나머지에는 '검은색'이 칠해져 있다고 하자. 그리고 아무도 모르게 A, B 항아리에서 각각 반쪽 공들을 하나씩 집어내어 서로 붙여 완전한 공을 만든 후 C항아리 속에 모두 넣어두었다고 하자. 그러면 C항아리 속에 있는 공들의 숫자(X)와 색깔(Y)에 대한 결합확률분포는 아래의 표와 같이 정리될 수 있다. 여기서 우리는 p만 미지의 모수이고 나머지 α와 β는 알려진 값으로 가정하는데, 이는 X와 Y의 결합확률분포 그 자체는 불확정적이지만 이의 행 주변분포와 열 주변분포는 완전히 알려져 있음을 의미한다.

[모집단]

색깔 \ 숫자	흰색	검은색	계
1	p	$\alpha - p$	α
2	$\beta - p$	$1 - \alpha - \beta + p$	$1 - \alpha$
계	β	$1 - \beta$	1

결합된 공들이 들어 있는 C항아리로부터 n개의 공을 비복원 임의추출하여 숫자와 색깔을 조사한 결과 다음과 같은 표본자료가 얻어졌다고 하자.

[표본]

색깔 \ 숫자	흰색	검은색	계
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
계	$n_{.1}$	$n_{.2}$	n

α, β의 값을 알고 있는 상황에서 우리는 주어진 표본으로부터 '공에 적힌 숫자와 칠해진 색깔간에는 통계적으로 서로 독립이다'라는 가설을 어떻게 검정할 것인가? 아주 쉬운 것으로 여겨지고 있는 이 문제에 대한 전통적인 해법은 이 문제를 적합성(goodness-of-fit)의 문제로 인식하고 다음과 같은 검정통계량(X^2 또는 G^2)을 사용하는 것이다.

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}, \quad G^2 = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right). \tag{2.1}$$

여기서, μ_{ij}는 귀무가설이 사실일 때 n_{ij}의 기대값으로, 본 문제에서는 귀무가설이 $H_0 : p = \alpha\beta$로 표현되어질 수 있기 때문에 다항분포의 기대값 공식에 의해 $\mu_{11} = n\alpha\beta$, $\mu_{12} = n\alpha(1-\beta)$, $\mu_{21} = n(1-\alpha)\beta$, $\mu_{22} = n(1-\alpha)(1-\beta)$가 된다. 한편 대표본 이론에 의하면 제기된 문제의 경우 검정통계량 X^2와 G^2는 독립성 가설하에서 자유도가 3인 카이제곱 분포에 근사하게 되며(Agresti, 1996), 그 값이 유의할 정도로 크면 독립이라는 가설을 기각하게 된다.

본 연구는 위에서 제기된 '공의 숫자와 색깔' 문제 상황과 같이 행과 열의 주변분포들이 이미 정확히 알려진 경우 두 변수간의 독립성 가설을 검정하고자 할 때 이를 적합성의 문제로만 인식하고 무보정의 자료에 대해 식(2.1)의 X^2 또는 G^2을 적용하는 전통적인 카이제곱 검정법이 과연 타당한 것인지에 관해 몇 가지 점에서 의문을 가지는 것으로부터 출발되었다.

첫 번째의 의문은 제기된 문제에서의 귀무가설의 수식적 표현과 논리적 의미에 관한 것이다. 귀무가설을 수식적인 형태로 표현하면 $H_0 : p = \alpha\beta$가 됨은 자명한 사실이지만, 이 귀무가설의 논리적 의미가 아래 두 가지 중 과연 어느 것인가 하는 것이 문제가 된다.

$$H_0 : [\text{숫자와 색깔은 독립}] \text{ and } [\text{주변분포들에 관한 주어진 정보가 사실}]$$
$$H_0 : [\text{숫자와 색깔은 독립}], \text{ given } [\text{주변분포들에 관한 주어진 정보가 사실}]$$

왜냐하면 그 차이에 따라 귀무가설이 기각되지 못하는 경우에서도 귀무가설의 해석에 차이가 있지만 귀무가설이 기각되었을 때 수용하게 되는 대립가설인 H_1 : *not* H_0의 논리적 의미와 해석도 당연히 달라지게 되기 때문이다. 그 이유는 논리학에 있어서의 아주 간단한 항진명제들인 '$not(\Theta_1\ and\ \Theta_2) = not(\Theta_1)\ or\ not(\Theta_2)$'와 '$not(\Theta_1,\ given\ \Theta_2) = not(\Theta_1),\ given\ \Theta_2$'을 생각해 본다면 쉽게 이해가 될 수 있다. 사실 제기된 문제의 상황적 입장과 얻고자 하는 결론을 고려한다면 귀무가설과 대립가설의 논리적 의미는 당연히 다음과 같아야 옳을 것이다.

$$H_0 : [\text{숫자와 색깔은 독립}], \text{ given } [\text{주변분포들에 관한 주어진 정보가 사실}]$$
$$H_1 : [\text{숫자와 색깔은 종속}], \text{ given } [\text{주변분포들에 관한 주어진 정보가 사실}]$$

즉 이는 '주변분포에 관한 주어진 정보가 사실'이라는 것을 대전제로, '숫자와 색깔은 독립'이라는 것을 소전제로 본 입장이다. 이러해야 함에도 불구하고 위의 전통적인 카이제곱 검정통계량들은 ', given'의 의미마저도 'and'의 의미로 해석하여 사용된다는 점에서 '공의 숫자와 색깔' 상황에서의 검정통계량으로 사용되는 데에는 문제가 있다고 할 수 있다. 즉 이들은 다음과 같은 가설들에 대한 검정통계량으로서는 적합한 것이라고 할 수 있다.

$$H_0 : [\text{숫자와 색깔은 독립}] \text{ and } [\text{주변분포들에 관한 주어진 정보가 사실}]$$
$$H_1 : [\text{숫자와 색깔은 종속}], \text{ or } [\text{주변분포들에 관한 주어진 정보가 거짓}]$$

그러면 과연 제기된 본 문제에 적합한 검정통계량은 무엇이어야 할 것인가?

두 번째 의문은 다음 표의 경우처럼 표본의 주변분포들이 이미 정확히 알려진 모집단 주변분포들과 상당한 차이를 보이고 있는 경우, 즉 대전제 하에서 아주 드물게 발생될 수 있

는 그런 표본이 얻어진 경우에서도 독립성 가설(소전제)에 대한 검정을 함에 있어서 표본에 대한 아무런 보정의 절차도 없이 그냥 전통적인 카이제곱 검정통계량들(X^2 또는 G^2)을 이용하여 검정을 수행해도 아무런 문제가 없을까 하는 점이다.

[모집단 분포]

색깔 \ 숫자	흰색	검은색	계
1	p	$0.7-p$	0.7
2	$0.4-p$	$p-0.1$	0.3
계	0.4	0.6	1.0

[표본 분할표]

색깔 \ 숫자	흰색	검은색	계
1	18	12	30
2	42	28	70
계	60	40	100

위의 표본은 가상 예로서 지극히 극단적인 경우라고 할 수 있지만, 어떤 표본추출 방법을 사용하든, 비록 이 정도는 아니더라도, 알려진 모집단 정보와 어느 정도의 크고 작은 차이를 보이는 표본이 얻어질 수 있는 가능성은 표본추출이라는 메카니즘에 의해 항상 상존하며, 실제로 큰 차이를 보이는 경우도 종종 발생되고 있다. 그 차이가 크든 작든 이는 분명히 카이제곱 검정통계량의 값을 크게 하는데 기여함은 분명하며, 이는 표본의 편향성이 검정의 결과에 크게 영향을 줄 수도 있음을 의미한다. 따라서 모집단에 관한 알려진 정보와 표본으로부터 얻어진 정보와의 차이가 존재하는 경우 통계적 추론의 과정에서 주어진 모집단 정보를 이용하여 그 차이를 전통적인 카이제곱 검정통계량의 값에서 보정할 수 있는 어떤 조치가 취해져야 하는 것은 당연하다고 생각할 수 있다. 문제는 그 보정법이 무엇인가 하는 점이다. 본 연구에서는 주어진 정보하에서의 최우추정 표본의 사용을 주장한다.

셋째, 피어슨 타입의 검정통계량(X^2)이나 우도비 검정통계량(G^2)은 모두 원 모형(underlying model)과 귀무가설을 구분하지 않고 귀무가설모형이라는 것에 통합하여 설명하고 있다는 점이 다른 일반적인 통계적 이론들과 일치하지 않는다는 점이다. 참고로 $N(\mu,\sigma_0^2)$로부터 얻어진 임의표본 $X_1, X_2, \cdots, X_n$을 이용하여 귀무가설 $H_0 : \mu = \mu_0$을 검정하는 문제를 생각해 보자. 이 경우 원 모형과 귀무가설은 다음과 같이 분명히 구분되어 표현되어지고 있다.

$$\text{원 모 형} : X_i = \mu + \epsilon_i \text{ for all } i, \epsilon_i \sim N(0,\sigma_0^2) \text{ for all } i, \epsilon_i \perp \epsilon_j \text{ for all } i,j(i \neq j)$$

$$\text{귀무가설} : \mu = \mu_0$$

사실 이 문제에 대한 가설검정에서 우리는 원 자료와 귀무가설 간의 차이라고 볼 수 있는 $\sum_i (X_i-\mu_0)^2$을 검정통계량으로 이용하는 것이 아니라, 원 모형하에서 검정대상의 모수인 μ를 추정량 $\bar{X}$로 추정한 후 이와 귀무가설 간의 차이라고 볼 수 있는 $n(\bar{X}-\mu_0)^2$을 검정통계량으로 이용한다. 이렇게 하는 이유는 여러 가지로 설명될 수 있지만 원 자료에 내재된 변이성에 해당되는 몫인 $\sum_i (X_i-\bar{X})^2$만큼을 원 자료와 귀무가설 간의 차이부분에서 제거한다는 의미도 지니고 있다. 참고로, 원 자료와 귀무가설 간의 차이부분은 원 자료와 추정량 간의 차이부분과 추정량과 귀무가설 간의 차이부분의 합으로 분리될 수 있음은 아래의 수식을 통해 쉽게 이해될 수 있을 것이다.

$$\sum_i (X_i-\mu_0)^2 = \sum_i (X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2.$$

이제 위에서 제기한 '공의 숫자와 색깔' 문제에 대해 원 모형과 귀무가설을 구분하여 표현해 보면 다음과 같다.

$$\text{원 모 형} : (n_{11}, n_{12}, n_{21}, n_{22}) \sim Multinomial(n, p, \alpha - p, \beta - p, 1 - \alpha - \beta + p)$$
$$\text{귀무가설} : p = \alpha\beta$$

이 문제에 대한 검정에서도 p의 추정량과 귀무가설 간의 차이를 이용하여 검정하여야 할 것이지만 피어슨의 검정통계량(X^2)이나 우도비 검정통계량(G^2) 모두 p의 추정량과 귀무가설 간의 차이가 아닌 원 자료와 귀무가설 간의 차이를 이용하여 검정하고 있다는 점이 이들 검정통계량들의 적절성과 타당성에 대한 회의감을 더하게 한다.

이러한 의문들을 바탕으로 본 연구에서는 모집단에 관한 주변분포들이 이미 알려진 상황에서, 얻어진 자료가 모집단 정보로부터 편향되어 있든 그렇지 않든 관계없이, 귀무가설이 기각되지 못하는 경우라면 '표본이 얻어진 모집단에서는 두 범주형 변수들이 통계적으로 독립이다', 기각되는 경우라면 '표본이 얻어진 모집단에서는 두 범주형 변수들은 통계적으로 독립이라고 볼 수 없다'라는 통계적 결론을 내릴 수 있는 검정방법을 유도한다. 이는 원자료 분할표를 주어진 모집단 부분정보를 활용하여 보정한 후 보정된 분할표에 대해 카이제곱 검정법을 적용하는 방법이라고 할 수 있다. 물론 보정에 따른 자유도의 차이는 존재한다.

3. 자료보정 카이제곱 독립성 검정법

범주형인 두 확률변수 X와 Y의 독립성을 검정하기 위해 단일 모집단으로부터 n개의 독립표본을 얻고, 이를 조사·정리하여 얻은 크기 $I \times J$인 2차원 분할표 자료를 $\{n_{ij}\}$라고 하자. 본 연구에서는, 2절에서 서술된 의문들을 근거로 하여, 두 확률변수 X와 Y의 독립성을 검정하고자 할 때 만일 모집단의 행 주변분포 $\{\alpha_i\}$와 열 주변분포 $\{\beta_j\}$가 완전히 알려진 경우라면 2절에서 제시된 전통적인 X^2 또는 G^2 검정통계량이 아닌 다음의 검정통계량들이 사용되어야 한다고 제안한다.

$$\tilde{X}^2 = \sum_{i,j} \frac{(m_{ij} - n\alpha_i\beta_j)^2}{n\alpha_i\beta_j}, \quad \tilde{G}^2 = 2\sum_{i,j} n_{ij} \log\left(\frac{m_{ij}}{n\alpha_i\beta_j}\right). \tag{3.1}$$

여기서 X^2 또는 G^2 검정통계량들과의 차이와 관련하여 유의할 점은 m_{ij}는 (i,j)번째 칸의 표본빈도인 n_{ij}가 아니라 i번째 행 확률들의 합이 α_i이고 j번째 열 확률들의 합이 β_j인 다항분포 $Multinomial(n, p_{11}, \cdots, p_{1J}, \cdots, p_{I1}, \cdots, p_{IJ}; \sum_j p_{ij} = \alpha, \sum_i p_{ij} = \beta_j, \ for\ all\ i,j)$ 하에서 최우추정법으로 추정된 (i,j)번째 칸의 추정빈도($= m_{ij}$)라는 것이다.

사실 검정통계량 $\tilde{G}^2$은 주어진 상황에 해당되는 우도비 검정통계량이고, $\tilde{X}^2$는 $\tilde{G}^2$에 대응되는 피어슨 타입의 검정통계량인데 이 검정통계량들은 귀무가설 하에서 자유도가 $df = (I-1)(J-1)$인 카이제곱 분포에 근사적으로 따르게 된다. 이의 수리적 증명과정을 기술하면 다음과 같다.

우선 $(N_{11}, N_{12}, \cdots, N_{IJ})$가 $Multinomial(n, p_{11}, p_{12}, \cdots, p_{IJ}; \sum_j p_{ij} = \alpha_i, \sum_i p_{ij} = \beta_j,$ for all $i,j)$인 다항분포로부터 얻어진 확률표본이라고 하자. 여기서 모든 α_i와 β_j는 알려진

값으로, $\sum_i \alpha_i = 1$, $\sum_j \beta_j = 1$을 만족하는 것으로 한다. 이 때 독립성 귀무가설은 $H_0 : p_{ij} = \alpha_i \beta_j$ for all i, j로 표현할 수 있고, 대립가설은 $H_1 : not\ H_0$가 되는데 이에 대한 우도비 검정법은 다음과 같이 유도될 수 있다.

전체 모수공간 Ω_M과 귀무가설하에서의 모수공간 Ω_N은 아래와 같으며, 그 차원은 $\dim(\Omega_M) = (I-1)(J-1)$, $\dim(\Omega_N) = 0$임을 쉽게 알 수 있다.

$$\Omega_M = \{(p_{11}, p_{12}, \cdots, p_{IJ}) | p_{ij} \geq 0 \text{ for all } i,j, \sum_j p_{ij} = \alpha_i \text{ for all } i, \sum_i p_{ij} = \beta_j \text{ for all } j\}$$

$$\Omega_N = \{(p_{11}, p_{12}, \cdots, p_{IJ}) | p_{ij} = \alpha_i \beta_j \text{ for all } i,j\}$$

그리고 N_{ij}의 결합확률밀도함수를 다음과 같이 표현한다면

$$f(n_{ij}) = f(n_{11}, n_{12}, \cdots, n_{IJ}) = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} p_{ij}^{n_{ij}},$$

이로부터 $Sup_{\Omega_M} f(n_{ij}) = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \left\{\frac{m_{ij}}{n}\right\}^{n_{ij}}$, $Sup_{\Omega_N} f(n_{ij}) = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} (\alpha_i \beta_j)^{n_{ij}}$ 을 얻을 수 있다. 여기서 m_{ij}/n은 Ω_M공간에서 p_{ij}에 대한 최우추정량을 나타낸다. 따라서 일반화 우도비(generalized likelihood-ratio) Λ와 $-2\log\Lambda$는 다음과 같이 정리된다.

$$\Lambda = \prod_{i,j} \left\{\frac{m_{ij}}{n\alpha_i\beta_j}\right\}^{-n_{ij}}, \quad -2\log\Lambda = 2\sum_{i,j} n_{ij} \log\left\{\frac{m_{ij}}{n\alpha_i\beta_j}\right\}.$$

그런데 일반적으로 널리 알려진 '일반화 우도비의 점근분포 정리(Mood, et al. 1974. 440쪽 참조)'에 의하면 대표본의 경우 $-2\log\Lambda$는 근사적으로 카이제곱분포를 따름을 알 수 있다. 물론 이 때의 자유도는 $\dim(\Omega_M) - \dim(\Omega_N) = (I-1)(J-1)$가 된다.

참고로, 실제 자료분석에서 위의 $\tilde{X}^2$ 또는 $\tilde{G}^2$ 검정통계량의 값을 계산하기 위해서는 m_{ij}라는 최우추정값들을 도출하는 계산과정이 약간 필요하지만 최적(optimization)문제를 푸는 좋은 컴퓨터 프로그램들(예를 들어, SAS/OR의 Proc NLP)이 아주 많이 개발·보급되어 있기 때문에 큰 어려움이 없이 그 값들을 구할 수 있다.

4. 전통적 무보정 검정법과 제안된 보정 검정법의 검정력 비교

본 연구에서는 모집단의 부분정보가 주어진 경우에서 이 정보를 활용한 카이제곱 독립성 검정법으로서 2장 식(2.1)의 검정통계량을 사용하는 전통적인 카이제곱 검정법을 편의상 '무보정 검정법', 3장 식(3.1)의 검정통계량을 사용하는 카이제곱 검정법을 '보정 검정법'이라 약칭한다. 본 4장에서는 컴퓨터 모의실험을 통하여 이 두 검정법의 검정력을 비교하고자 한다.

모의실험 방법은 다음과 같다. 행주변분포가 $(\alpha, 1-\alpha)$, 열주변분포가 $(\beta, 1-\beta)$인 것으로 완전히 알려진 2×2 분할표 분포들 중 다음 표 4.1과 같은 형태의 분포들을 대상으로 하였다. 여기에서 δ의 값이 0이면 독립성을 만족하는 분포 즉 귀무분포가 되고, δ값이 클수록 귀무분포로부터 더 먼 대립분포가 된다. 본 모의실험에서는 $\alpha, \beta = .5, .6, .7$의 각 경우

에 대하여 귀무분포($\delta = 0$)와 두 개의 대립분포($\delta = .05, .10$)를 고려하였다. 그리고 각 상황에서 획득한 표본의 수(n)는 각 칸의 기대도수가 최소 5이상이 되도록 하였다($\alpha = \beta = .5$, $\delta = .10$인 경우만 제외). 예를 들어, $\alpha = \beta = .6$, $\delta = .10$인 경우 칸 확률이 가장 작은 값은 0.14이므로 표본의 수가 40이면 이 칸의 기대도수는 5.6이 된다.

표 4.1: 모의실험에 사용된 분포들의 형태

$X \backslash Y$	C_1	C_2	계
R_1	$\alpha\beta + \delta$	$\alpha(1-\beta) - \delta$	α
R_2	$(1-\alpha)\beta - \delta$	$(1-\alpha)(1-\beta) + \delta$	$1-\alpha$
계	β	$1-\beta$	1

모의실험에는 SAS 8.2 버전이 이용되었으며, 모의실험용으로 작성된 프로그램은 부록A에 별첨하였다. 본 연구의 모의실험의 결과를 재확인할 필요성이 있는 경우를 위하여 표 4.2의 모의실험 결과표에 난수발생에 필요한 초기값(seed)도 표시하였다. 각 경우에서의 반복회수(iteration)는 모두 5,000번씩으로 하였으며, 1종 오류율(유의수준)은 5%인 경우로 고정하였다.

표 4.2를 보는 법을 표의 첫 행을 예로 하여 설명하면 다음과 같다. α, β의 값이 0.5, δ의 값이 0, n의 값이 30이므로 $\alpha = \beta = 0.5$, $\delta = 0$에 해당되는 표 4.1의 모집단에서 표본크기가 30인 표본을 랜덤하게 추출하였다는 뜻이다. 이 때 난수생성에 적용한 초기값(seed)은 54321이다. 이 표본에 대해 X^2, G^2, $\tilde{X}^2$, $\tilde{G}^2$의 네 가지 검정통계량을 계산하여 '독립'이라는 귀무가설의 기각여부를 유의수준 5%하에서 검정한다. 이 때 적용된 귀무분포는 X^2, G^2 검정통계량에 대해서는 자유도가 3인 카이제곱분포를 $\tilde{X}^2$, $\tilde{G}^2$ 검정통계량에 대해서는 자유도가 1인 카이제곱분포이다. 이런 과정을 5,000번 반복한 결과 X^2 검정통계량의 값들 중 5.36%, G^2 검정통계량의 값들 중 5.52%, $\tilde{X}^2$ 검정통계량의 값들 중 4.56%, $\tilde{G}^2$ 검정통계량의 값들 중 4.56%가 기각역을 초과한 것으로 나타났음을 의미한다. 표 4.3은 주어진 표본크기 하에서 δ의 변화에 따른 검정력의 변화를 쉽게 알아볼 수 있도록 표 4.2를 재구성한 것에 불과하다.

표 4.2에서 $\delta = 0$인 경우는 귀무가설의 상황들에 해당되는데 이의 총 12가지 각각에서 5%로 설정된 1종의 오류가 유지되는지를 살펴볼 수 있다. 네 가지의 검정법 모두 대체로 만족할만한 수준으로 나타났으며, 12가지 귀무가설 상황들의 평균 1종 오류율을 구해보면 X^2, G^2, $\tilde{X}^2$, $\tilde{G}^2$ 각각이 4.91, 5.25, 4.99, 5.11로 피어슨 타입의 보정 검정법이 설정된 1종 오류율 5%를 가장 잘 유지하는 것으로 나타났다.

$\delta = 0.05$, $\delta = 0.1$은 대립가설의 상황으로 δ의 절대값이 클수록 '독립'이라는 귀무가설로부터 더 먼 대립가설이라고 할 수 있다. 표 4.2를 통하여 모든 대립가설의 상황에서 표본이 커질수록 검정력이 높아져야 한다는 당연한 사실을 확인할 수 있다. 표 4.3을 통해서는 주어진 표본크기 하에서는 δ값이 커짐에 따라 검정력도 커짐을 확인할 수 있다. 그러나 중요한 것은 무보정의 경우보다는 보정의 경우가 검정력이 훨씬 높다는 것을 확인할 수 있다.

예를 들면 표 4.3에서 $\alpha = \beta = 0.5$, $n = 30$의 경우 귀무가설하에서는 보정법이 무보정법보다 기각율이 더 낮음에도 불구하고 δ의 값이 0.05, 0.10으로 커짐에 따라 무보정의 경우는 13%, 43% 정도이지만 보정의 경우는 17%, 59%로 무보정의 경우보다 훨씬 큼을 알 수 있다. 이러한 현상이 모의실험된 모든 대립가설 상황에서도 재현되고 있음을 통해 보정 검정법이 무보정 검정법보다 더 우수함을 확인할 수 있었다.

표 4.2: 모의실험 결과

α, β	δ	n	seed	무보정		보정	
				X^2	G^2	$\tilde{X}^2$	$\tilde{G}^2$
0.5	0	30	54321	5.36	5.52	4.56	4.56
		50	54322	5.12	5.12	6.40	6.40
		70	54323	4.48	4.64	3.94	3.94
		90	54324	5.46	5.68	5.04	5.04
0.6		40	54325	4.50	5.28	4.40	4.60
		70	54326	4.60	5.04	4.82	5.16
		100	54327	4.68	5.06	4.80	4.86
		130	54328	5.40	5.88	5.00	5.06
0.6		60	54329	4.58	5.08	5.70	5.64
		100	54330	5.04	5.32	5.52	6.02
		140	54331	5.22	5.50	4.94	5.12
		180	54332	4.42	4.92	4.74	4.92
평 균				4.91	5.25	4.99	5.11
0.5	0.05	30	54333	13.44	13.46	17.60	17.60
		50	54334	18.56	18.50	33.90	33.90
		70	54335	24.60	25.18	34.92	34.92
		90	54336	32.06	32.40	45.66	45.66
0.6		40	54337	17.28	18.26	26.26	27.02
		70	54338	27.66	27.82	41.38	41.76
		100	54339	38.24	38.64	55.32	55.28
		130	54340	48.72	49.06	65.42	65.58
0.7		60	54341	29.96	29.34	46.26	45.34
		100	54342	49.08	48.24	65.68	64.52
		140	54343	65.04	64.06	80.14	79.48
		180	54344	76.20	75.58	88.32	87.58
0.5	0.10	30	54345	42.96	43.84	59.18	59.18
		50	54346	67.16	67.38	86.38	86.38
		70	54347	83.30	83.46	91.90	91.90
		90	54348	92.32	92.64	97.22	97.22
0.6		40	54349	58.64	60.74	76.66	77.04
		70	54350	86.90	87.68	95.32	95.12
		100	54351	96.14	96.34	99.28	99.24
		130	54352	98.92	98.94	99.80	99.80
0.7		60	54353	87.96	87.86	95.94	95.56
		100	54354	98.66	98.80	99.76	99.72
		140	54355	99.98	99.98	100.00	100.00
		180	54356	100.00	100.00	100.00	100.00

표 4.3: 모의실험 결과(표 4.2를 재배치 한 것)

α, β	δ	n	seed	무보정		보정	
				X^2	G^2	$\tilde{X}^2$	$\tilde{G}^2$
0.5	0	30	54321	5.36	5.52	4.56	4.56
	0.05		54333	13.44	13.46	17.60	17.60
	0.10		54345	42.96	43.84	59.18	59.18
	0	50	54322	5.12	5.12	6.40	6.40
	0.05		54334	18.56	18.50	33.90	33.90
	0.10		54346	67.16	67.38	86.38	86.38
	0	70	54323	4.48	4.64	3.94	3.94
	0.05		54335	24.60	25.18	34.92	34.92
	0.10		54347	83.30	83.46	91.90	91.90
	0	90	54324	5.46	5.68	5.04	5.04
	0.05		54336	32.06	32.40	45.66	45.66
	0.10		54348	92.32	92.64	97.22	97.22
0.6	0	40	54325	4.50	5.28	4.40	4.60
	0.05		54337	17.28	18.26	26.26	27.02
	0.10		54349	58.64	60.74	76.66	77.04
	0	70	54326	4.60	5.04	4.82	5.16
	0.05		54338	27.66	27.82	41.38	41.76
	0.10		54350	86.90	87.68	95.32	95.12
	0	100	54327	4.68	5.06	4.80	4.86
	0.05		54339	38.24	38.64	55.32	55.28
	0.10		54351	96.14	96.34	99.28	99.24
	0	130	54328	5.40	5.88	5.00	5.06
	0.05		54340	48.72	49.06	65.42	65.58
	0.10		54352	98.92	98.94	99.80	99.80
0.7	0	60	54329	4.58	5.08	5.70	5.64
	0.05		54341	29.96	29.34	46.26	45.34
	0.10		54353	87.96	87.86	95.94	95.56
	0	100	54330	5.04	5.32	5.52	6.02
	0.05		54342	49.08	48.24	65.68	64.52
	0.10		54354	98.66	98.80	99.76	99.72
	0	140	54331	5.22	5.50	4.94	5.12
	0.05		54343	65.04	64.06	80.14	79.48
	0.10		54355	99.98	99.98	100.00	100.00
	0	180	54332	4.42	4.92	4.74	4.92
	0.05		54344	76.20	75.58	88.32	87.58
	0.10		54356	100.00	100.00	100.00	100.00

5. 분포공간, 차원, 그리고 기하

확률변수 X와 Y는 각각 I개와 J개의 범주를 가지는 범주형 확률변수로서, (X, Y)가 2차원 분할표의 (i, j)칸에 떨어질 확률을 p_{ij}라고 하자. 그리고 전체 표본의 크기를 n, 이들

중 (i,j)칸에 해당되는 표본빈도를 n_{ij}, i번째 행 표본빈도들의 합을 $n_{i.}$, j번째 열 표본빈도들의 합을 $n_{.j}$로 나타내기로 하자.

기하적인 입장에서 크기가 $I \times J$인 2차원 분할표 형태의 모든 다항분포는 IJ차원 공간에 존재하는 $(IJ-1)$차원 심플렉스(simplex)인 Ω_S상의 한 점으로 나타내어질 수 있기 때문에 이 심플렉스는 모수에 대한 제약조건이 없는 다항분포의 공간이라고 할 수 있으며, 얻어진 분할표 자료에 대응되는 표본 다항분포는 이 Ω_S상의 한 점에 대응하게 된다. 그리고 행 주변분포 $\{\alpha_i\}$와 열 주변분포 $\{\beta_j\}$가 모두 알려진 다항분포의 공간을 Ω_M, 독립성 조건만을 만족하는 다항분포의 공간을 Ω_H, Ω_M과 Ω_H의 공통집합인 공간 즉 행과 열 주변분포들이 모두 알려져 있으면서 독립성 조건도 만족하는 다항분포의 공간을 Ω_N으로 나타내기로 한다면 이들 공간들과 그 차원들을 수학적으로 기술하면 다음과 같다.

$$
\begin{aligned}
\Omega_S &= \{(p_{11}p_{12}\cdots p_{IJ}) | p_{ij} \geq 0 \text{ for all } i,j, \sum_{i,j} p_{ij} = 1\} \quad \dim(\Omega_S) = IJ-1 \\
\Omega_M &= \{(p_{11}p_{12}\cdots p_{IJ}) | p_{ij} \geq 0 \text{ for all } i,j, \\
&\qquad \sum_j p_{ij} = \alpha_i \text{ for all } i, \sum_i p_{ij} = \beta_j \text{ for all } j\} \quad \dim(\Omega_M) = (I-1)(J-1) \\
\Omega_H &= \{(p_{11}p_{12}\cdots p_{IJ}) | p_{ij} \geq 0 \text{ for all } i,j, \\
&\qquad \sum_{i,j} p_{ij} = 1, p_{ij} = p_{i.}p_{.j} \text{ for all } i,j\} \qquad \dim(\Omega_H) = I+J-2 \\
\Omega_N &= \{(p_{11}p_{12}\cdots p_{IJ}) | p_{ij} = \alpha_i\beta_j \text{ for all } i,j\} \qquad \dim(\Omega_N) = 0
\end{aligned}
$$

이들 분포공간들 간에는 $\Omega_N \subset \Omega_M \subset \Omega_S$, $\Omega_N \subset \Omega_H \subset \Omega_S$의 포함관계가 있음을 쉽게 확인할 수 있다. 그리고 다음의 그림 5.1은 이 공간들에 대한 설명과 이해를 위해 편의적으로 작성된 것으로, 분할표와 관련한 기하의 상세한 표현과 서술 그리고 관계들은 Fienberg(1968)과 Fienberg & Gilbert(1970)을 참조할 수 있겠다.

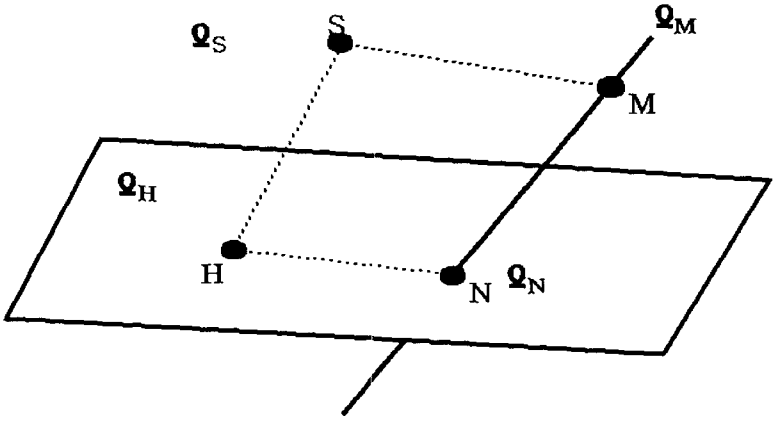

그림 5.1: 2차원 분할표 형태의 다항분포와 관련한 분포공간들

크기가 $I \times J$인 2차원 분할표 자료에 대하여 기저분포(underlying distribution)로서 다항분포를 가정했을 때의 우도함수는 $L(\underline{p}; n_{ij}) = n! \prod_{ij}(p_{ij}^{n_{ij}}/n_{ij}!)$이 되는데, 이 우도함수의 값이 Ω_S, Ω_M, Ω_H, Ω_N의 각 공간내에서 최대가 되는 위치를 그림 5.1에서는 각각 S, M, H, N 점으로 나타내었다. 만일 m_{ij}를 np_{ij}에 대한 Ω_M상에서의 최우추정량이라 한다면, 점 S는 표본 다항분포 $\{n_{ij}/n\}$, 점 M은 주변분포 조건을 만족하는 최우추정 다항분포 $\{m_{ij}/n\}$, 점 H는 독립성 조건을 만족하는 최우추정 다항분포 $\{n_{i.}n_{.j}/n^2\}$, 점 N은 주변분포 조건과 독립성 조건을 모두 만족하는 다항분포 $\{\alpha_i\beta_j\}$에 각각 해당된다고 할 수 있다.

여기에서 우리는 $\Omega_A \subset \Omega_B$ 조건을 만족하는 두 공간에 대하여 공간 Ω_A에서 공간 Ω_B까지의 거리를 $\overline{AB}$로 표현하고, 이를 $\overline{AB} = -2 \times \ln(\max_{\underline{p}\in\Omega_A} L(\underline{p}; n_{ij}) / \max_{\underline{p}\in\Omega_B} L(\underline{p}; n_{ij}))$로 정의한다면, $\overline{SM}$, $\overline{SH}$, $\overline{SN}$, $\overline{MN}$, $\overline{HN}$은 각각 다음과 같이 쉽게 재표현 되어질 수 있다.

$$\overline{SM} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{m_{ij}}\right) \qquad \overline{SH} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{n_{i.}n_{.j}/n}\right)$$

$$\overline{SN} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{n\alpha_i\beta_j}\right) \qquad \overline{MN} = 2\sum_{i,j} n_{ij} \log\left(\frac{m_{ij}}{n\alpha_i\beta_j}\right)$$

$$\overline{HN} = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{i.}n_{.j}/n}{n\alpha_i\beta_j}\right)$$

우리는 이 거리정의와 재표현들로부터, 그리고 Bickel & Doksum(1977)에서도 설명되어 있는 일반우도비근사성(General likelihood ratio approximation)에 의해 다음과 같은 사실들을 모두 쉽게 얻을 수 있다.

첫째(분할성), $\overline{SN} = \overline{SM} + \overline{MN}$, $\overline{SN} = \overline{SH} + \overline{HN}$

둘째(점근적 분포성), $\overline{SM} \dot{\sim} X^2(df = I + J - 2)$, $\overline{SH} \dot{\sim} X^2(df = (I-1)(J-1))$

$\overline{SN} \dot{\sim} X^2(df = IJ - 1)$, $\overline{MN} \dot{\sim} X^2(df = (I-1)(J-1))$,

$\overline{HN} \dot{\sim} X^2(df = I + J - 2)$.

여기서, ' $\dot{\sim}$ '는 점근적으로 분포가 동일함(asymptotically distributional equivalence)을 나타낸다.

셋째(독립성), $\overline{SM} \perp \overline{MN}$, $\overline{SH} \perp \overline{HN}$

6. 관련 가설들에 따른 검정통계량들

5장에서 기하적 입장에서 살펴본 거리들 즉 우도비 검정통계량들은 제각각 관련 가설상황에서의 검정통계량으로 그 역할을 다 할 수 있다. 이들을 정리하기 위해 다음과 같은 네 가지의 검정대상 가설상황들을 생각해 보자.

가설상황1) $H1_0$: 주어진 행, 열 주변분포들이 모두 사실이다. *vs.* $H1_A$: *not* $H1_0$

가설상황2) $H2_0$: 행변수와 열변수는 독립이다. *vs.* $H2_A$: *not* $H2_0$

가설상황3) $H3_0$: $H2_0$ *and* $H1_0$ *vs.* $H3_A$: *not* $H3_0$

가설상황4) $H4_0$: $H2_0$, *given* $H1_0$ *vs.* $H4_A$: *not* $H4_0$

각 가설상황에서의 우도비 검정통계량은 각각 $\overline{SM}$, $\overline{SH}$, $\overline{SN}$, $\overline{MN}$이 됨은 간단하게 증명될 수 있다. 그리고 이 우도비 검정통계량의 형태로부터 우리는 각 가설상황에 대한 피어슨 타입의 검정통계량을 'intuitive and heuristic'한 관점에서 다음과 같이 도출할 수 있다.

$$\begin{aligned}
\overline{SM} = G_1^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{m_{ij}}\right) &\quad \dot{\sim} \quad X_1^2 = \sum \frac{(n_{ij}-m_{ij})^2}{m_{ij}},\\
\overline{SH} = G_2^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{n_{i.}n_{.j}/n}\right) &\quad \dot{\sim} \quad X_2^2 = \sum \frac{(n_{ij}-n_{i.}n_{.j}/n)^2}{n_{i.}n_{.j}/n},\\
\overline{SN} = G_3^2 = 2\sum n_{ij}\log\left(\frac{n_{ij}}{n\alpha_i\beta_j}\right) &\quad \dot{\sim} \quad X_3^2 = \sum \frac{(n_{ij}-n\alpha_i\beta_j)^2}{n\alpha_i\beta_j},\\
\overline{MN} = G_4^2 = 2\sum n_{ij}\log\left(\frac{m_{ij}}{n\alpha_i\beta_j}\right) &\quad \dot{\sim} \quad X_4^2 = \sum \frac{(m_{ij}-n\alpha_i\beta_j)^2}{n\alpha_i\beta_j}.
\end{aligned}$$

여기서, m_{ij}는 p_{ij}에 대한 Ω_M상에서의 최우추정량을, "$\dot{\sim}$"는 점근적 분포 동일성을 나타낸다.

참고로, 우도비 검정통계량들 간에는 $\overline{SN} = \overline{SM} + \overline{MN} = \overline{SH} + \overline{HN}$라는 관계식처럼 정확한 분할성이 성립되지만, 피어슨 타입의 검정통계량들 사이에는 아래와 같은 근사적 분할성만이 성립된다.

$$X_3^2 = X_1^2 + X_4^2 + \sum_{i,j} \frac{(n_{ij}^2 - m_{ij}^2)}{m_{ij}} \frac{(m_{ij}-n\alpha_i\beta_j)}{n\alpha_i\beta_j}.$$

참고문헌

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

[2] Alexander M. Mood, Franklin A. Graybill, and Duane C. Boes (1974). Introduction to The Theory of Statistics, 3rd Edition, Mcgraw-Hill.

[3] Fienberg, S. E. (1968). The Geometry of an Contingency Table. The Annals of Mathematical Statistics, Vol. 39, No. 4, 1186-1190.

[4] Fienberg, S. E. and Gilbert, J. P. (1970). The Geometry of a Two by Two Contingency Table. Journal of the American Statistical Association, Vol. 65, No. 330, 694-701.

[5] Peter J. Bickel and Kjell A. Doksum (1977). Mathematical Statistics-Basic Ideas and Selected Topics, Holden-day, Inc..

[2003년 6월 접수, 2003년 9월 채택]

부록 A: 모의실험용 SAS 프로그램

```
data full;
 input alpha beta delta n seed @@;
    p11=   alpha * beta + delta;   p12=   alpha *(1-beta) - delta;
    p21=(1-alpha)* beta - delta;   p22=(1-alpha)*(1-beta) + delta;
 iter=1;
 do while(iter <= 5000);
 /**** OBTAIN THE RANDOM SAMPLES ****/
    n11=0; n12=0; n21=0; n22=0;
    do i=1 to n;
       call rantbl(seed, p11, p12, p21, x);
            if x=1 then n11=n11+1; else if x=2 then n12=n12+1;
       else if x=3 then n21=n21+1; else             n22=n22+1;
    end;
    n1a=n11+n12; n2a=n21+n22; na1=n11+n21; na2=n12+n22;
 /**** DERIVATION OF THE M.L.E OF p11 ****/
    max=-10000000000;
    do x=max(0, alpha+beta-1+0.00001) to min(alpha-0.00001, beta-0.00001)
                                      by 0.00001;
       l=n11*log(x)+n12*log(alpha-x)+n21*log(beta-x)+n22*log(1-alpha-beta+x);
       if l > max then do max=l; mle_p11=x; end;
    end;
 /**** OBTAIN THE EXPECTED CONTINGENCY TABLE ****/
    C11=   alpha *beta *n;   C12=   alpha *(1-beta)*n;
    C21=(1-alpha)*beta *n;   C22=(1-alpha)*(1-beta)*n;
 /**** OBTAIN THE cell log CONTINGENCY TABLE ****/
    if n11=0 then l11=0 else l11 = n11*log(n11/C11);
    if n12=0 then l12=0 else l12 = n12*log(n12/C12);
    if n21=0 then l21=0 else l21 = n21*log(n21/C21);
    if n22=0 then l22=0 else l22 = n22*log(n22/C22);
 /**** OBTAIN THE MLE CONTINGENCY TABLE ****/
    m11=(     mle_p11)*n;   m12=(  alpha     -mle_p11)*n;
    m21=(beta-mle_p11)*n;   m22=(1-alpha-beta+mle_p11)*n;
    m1a=m11+m12; m2a=m21+m22; ma1=m11+m21; ma2=m12+m22;
 /**** H30: H10 and H20 vs. H31: not H30 ****/
    chisq3p= (n11-C11)**2/C11 + (n12-C12)**2/C12 + (n21-C21)**2/C21
                              + (n22-C22)**2/C22;
    chisq3l= 2*(l11 + l12 + l21 + l22);
    diff3=chisq3l-chisq3p;
    if(chisq3p>cinv(0.95,3)) then result3p='R' else result3p='A';
    if(chisq3l>cinv(0.95,3)) then result3l='R' else result3l='A';
```

```
 /**** H40: H20 given H10 vs H41: not H40 ****/
    chisq4p= (m11-C11)**2/C11 + (m12-C12)**2/C12 + (m21-C21)**2/C21
                              + (m22-C22)**2/C22;
    chisq4l=2*(n11*log(m11/C11) + n12*log(m12/C12) + n21*log(m21/C21)
                                + n22*log(m22/C22));
    diff4=chisq4l-chisq4p;
    if(chisq4p>cinv(0.95,1)) then result4p='R' else result4p='A';
    if(chisq4l>cinv(0.95,1)) then result4l='R' else result4l='A';
  output;
  iter=iter+1;
 end;
cards; 0.5 0.5 0.00 30 54321
;
proc means; var chisq3p chisq4p chisq3l chisq4l diff3-diff4; run;
proc corr;  var chisq3p chisq4p chisq3l chisq4l; run;
proc freq; tables result3p result3l result4p result4l / nocum; run;
```

Chi-Squared Test of Independence in Case that Two Marginal Distributions are Given Exactly *

Kwangjin Lee [1)]

ABSTRACT

If the given information is exact, though it is the little, we had better use it than not use in analysis. In this article, the problem of independence test in a contingency table is considered when two marginal distributions of a population are given exactly. For that case, a likelihood-ratio chi-squared test statistic and its Pearsonian type chi-squared test statistic are derived. By Monte Carlo Simulations the traditional chi-square tests and the derived tests are compared. And the related some testing problems are synthetically explained on a geometrical viewpoint.

Keywords: Contingency Table, Population Partial Information, Independence Test, Data Modification.

* The author wishes to acknowledge the financial support of Mokwon University, Korea, made in the program year of 1998.

1) Dept. of Statistics, Mokwon University, 302-729, Daejeon, Korea.
E-mail : leekj@mokwon.ac.kr