응용통계연구 제 14권 1호, 2001년, pp. 39-57

범주형 자료에서 경험적 베이지안 오분류 분석

임한승[1] 홍종선[2] 서문섭 [3]

요 약

범주형 자료에서 오분류는 자료를 수집하는 과정에서 발생될 수 있다. 오분류되어 있는 자료를 정확한 자료로 간주하여 분석한다면 추정결과에 편의가 발생하고 검정력이 약화되는 결과를 초래하게 되며, 정확하게 분류된 자료를 오분류라고 판단한다면 오분류의 수정을 위해 불필요한 비용과 시간을 낭비해야 할 것이다. 따라서 정확하게 분류된 표본인지 오분류된 표본인지를 판정하는 것은 자료를 분석하기 전에 이루어져야할 매우 중요한 과정이다. 본 논문은 $I \times J$분할표로 주어지는 범주형 자료에서 두 변수 중 하나의 변수에서만 오분류가 발생되는 경우에 오분류 여부를 검정하기 위해서 오분류 가능성이 없는 변수에 대한 주변합은 고정시키고, 오분류 가능성이 있는 변수의 주변합을 Sebastiani와 Ramoni(1997)가 제안한 Bound와 외부정보로 표현되는 Collapse의 개념, 그리고 베이지안 방법을 확장하여 자료에 적합한 모형과 사전정보를 고려한 사전모수를 다양하게 설정하면서 재분류하는 연구를 하였다. 오분류에 대한 정보를 얻기 위해서 Tenenbein(1970)에 의해 연구된 이중추출법을 이용하여 오분류 검정을 위한 새로운 통계량을 제안하였으며, 제안된 오분류 검정통계량에 관한 분포를 다양한 모의실험을 통하여 연구하였다.

주요용어: 결측자료, 경험적 베이지안 추정, 오분류, 사전모수, 사전정보, 사후추정량, 이중표본추출법, 외부정보, MC-검정통계량.

1. 서론

최근 여러 분야에서 다양한 종류의 대규모 여론조사가 시행되고 있으며, 조사방법도 설문지를 이용한 면접조사를 포함하여 다양한 방법을 사용하고 있다. 전화조사나 우편조사 혹은 인터넷을 이용한 여론조사는 시간과 비용을 줄일 수 있으며 동일한 비용으로 보다 많은 표본을 얻을 수 있다는 장점을 갖고 있다. 반면 이러한 저비용의 대규모 여론조사에서는 질문에 대해 무응답(nonresponse)나 결측값(missing data)이 자주 발생할 수 있으며, 응답되어진 질문이라도 까다로운 질문을 회피하려는 심리 때문에 피설문자에게 정확한 응답을 얻지 못하는 경우도 발생한다. 표본조사에는 오차가 발생하기 마련인데 발생원인에 따라 표본오차(sampling error)와 비표본오차(non-sampling error)로 구분된다. 비표본오차에는

1) (150-737) 서울 영등포구 여의도동 26-4 교보증권빌딩 8F, 한국기업평가, 책임연구원
E-mail: haslim@kmcc.com
2) (110-745) 서울 종로구 명륜동3가, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
3) 미국, Wright State University, Department of Mathematics and Statistics, 교수
E-mail: munsup.seoh@wright.edu

무응답오차(non-response error)와 응답오차(response error)로 나누어 고려할 수 있다. 무응답오차는 표본에 포함된 단위들에 대해 응답을 얻지 못한 경우에 발생하는 오차이며, 응답오차인 오분류(misclassification)는 자료를 수집하여 분류하는 과정에서 관측값이 해당 범주가 아닌 다른 범주에 잘못 분류되는 경우의 오류이다.

만약 오분류 되어있는 범주형 자료를 이용하여 분석하면 추정의 편의가 발생하며 검정의 검정력이 감소하게된다(Bross 1954). 따라서 수집된 자료의 분석에 앞서 오분류의 정도를 파악하는 것은 매우 중요하다. 범주형 자료에서의 오분류 문제를 해결하기 위해서 Bross(1954), Barron(1977), Hochberg(1977), Fuchs(1982) 등의 많은 학자들에 의해서 연구되어 왔는데 특히 Tenenbein(1970, 1971)은 주어진 조사비용하에서 오분류를 파악하기 위한 방법으로 비용을 고려한 이중추출법(double sampling scheme)을 제안하였으며 오분류 가능성이 있다고 생각되는 초기표본에서 초기표본보다 적은 수의 부표본을 뽑아 오분류의 정도를 파악하고 추정의 편의를 줄이려고 하였다.

범주형 자료에 대해 분할표에서 발생되는 무응답이나 결측값의 성질은 결측값 메카니즘(missing data mechanism)에 따라 나눌 수 있는데, 결측값 메카니즘이 MAR(missing at random) 또는 MCAR(missing complete at random)일 경우에는 무시할 수 있는 무응답(ignorable nonresponse: IN)으로 자료분석시 관측값만을 고려하게 된다. 반면 메카니즘이 MAR이 아니며 OAR(observed at random)도 아니라면 무시할 수 없는 무응답(nonignorable nonresponse: NIN)이며, 자료분석과정에서 무응답 자료를 무시하고 분석하면 표본의 임의성에 영향을 줄 수 있으며 추정결과에 편의가 발생한다(Kalton 1983). 최근에는 무시할 수 없는 무응답을 추정하기 위하여 베이지안 방법이 많이 이용되는데 Park과 Brown(1994)은 최대가능도 추정법(maximum likelihood estimation method)이 갖는 주변값 문제의 해결 위해 무응답이 발생한 칸에 사전확률분포를 가정하고 사후밀도함수를 구한 후 EM 알고리즘을 이용하여 사후밀도함수를 최대로 하는 추정값을 제안하였다. Sebastiani와 Ramoni(1997)는 베이지안 추정과정을 Bound와 Collapse단계로 나누어 추정량의 값의 범위와 외부정보로 표현되는 점추정량을 유도하였다. Tebaldi와 West(1998)는 Metropolis-Hasting 알고리즘과 같은 MCMC(Markov Chain Monte Carlo)기법을 이용하여 결측값을 추정하였다.

본 논문에서는 범주형 자료의 오분류 여부를 검정하기 위해서 무응답값을 추정하는 방법을 이용하였는데 $I \times J$분할표에서 오분류 가능성이 없는 변수에 대한 주변합은 고정시키고, 오분류 가능성이 있는 변수의 주변합을 무응답으로 고려하였으며 무응답값의 추정을 위해서 Sebastiani와 Ramoni(1997)가 제안한 Bound와 Collapse의 개념을 이용한 베이지안 추정법을 사용하였다. 추정을 위한 정보로는 Tenenbein(1970)에 의해 제안된 이중추출자료뿐만 아니라 표본조사비용(sampling cost)과 외부정보(external-information), 타 연구기관의 조사결과나 혹은 몇 년 전에 실시된 유사한 조사와 같은 사전정보(pre-information) 등을 고려하였다. 그리고 오분류의 정도파악을 위한 검정통계량인 MC-통계량(Misclassification statistic)을 제안하였으며 그 통계량의 분포를 연구하였다. 오분류 가능성이 있는 범주형 변수의 각 범주별로도 오분류 정도를 파악할 수 있는 범주별 검정통계량인 MC_i-통계량($i = 1, \ldots, I$)을 제안하였다. 이와 같은 연구는 다양한 모의실험을 이용하여 구한 오분류된 자료를 통하여 실시되었다. 모의실험은 각각 무시할 수 있는 무응답(IN)과 무시할 수 없는 무

응답(NIN)인 경우에 초기표본크기 , 이중추출율과 조건부확률에 대한 오분류율의 변화, 그리고 오분류 가능성이 없는 변수의 주변합의 크기와 조건부확률의 크기를 다양하게 변화시키면서 연구하였다.

2. 오분류 자료의 구조

일반적인 2차원 자료인 $I \times J$분할표는 범주형 변수 X와 Y로 구성되어 있으며 N개의 표본으로 이루어져 있다고 가정하자. (i,j)칸의 빈도수를 N_{ij} 그리고 i번째 행주변합과 j번째 열주변합을 각각 N_{i+}와 N_{+j}로 표기하자($N = N_{++}$). 주어진 2차원 분할표의 자료분석에 앞서 분할표의 오분류 여부를 파악하는 것은 매우 중요하다(Bross 1954). 행변수 X와 열변수 Y로 구성된 분할표에 대해서 변수별로 오분류가 있는 경우를 살펴보자. 우선 X의 범주에만 오분류가 발생할 경우와 Y의 범주에만 오분류가 있는 경우 그리고 X와 Y의 범주 모두에 오분류가 있을 경우를 고려할 수 있다. 본 논문에서 관심있게 연구되어지는 오분류된 자료의 형태는 두 변수(X와 Y)중의 한 변수(Y)의 범주에만 오분류의 가능성이 있으며 나머지 다른 변수(X)에는 오분류의 가능성이 없는 경우에 자료형태이다. 예를 들어 변수 X는 거주지역을 나타내는 변수이고 Y는 소득수준을 나타내는 변수라 하면 X의 범주는 오분류의 가능성이 희박할 것이며 반대로 Y의 범주에 대해서는 오분류의 가능성이 상대적으로 높을 것이다. 2×2분할표의 다른 예를 고려한다면 X는 성별을 나타내는 변수이고 Y는 특정 정당의 지지 여부를 나타내는 경우를 들 수 있다. 이와 같은 예제에서 변수 X의 주변합은 각 범주별로 정확하다고 고려되지만 변수 Y에 대해선 무응답의 경우와 같이 범주 의 구분이 명확하지 않다. 이와 같이 한 변수 Y의 범주에만 오분류된 경우에는 정확하다고 고려되는 변수 X의 주변합만은 정확한 자료의 정보로 간주한다.

본 논문에서는 범주형 자료의 오분류를 파악하기 위해서 이중추출법을 사용한다. Tenenbein(1970)이 제안한 이중추출법은 일변량의 경우로서 추출된 n ($< N$)개의 부표본 (subsample)을 N개의 초기표본 (initial sample)에서 부표본과 대응되는 n개의 표본과 교차하여 2×2분할표를 만들었다. 그러나 본 논문의 경우는 이변량의 경우를 고려하며, 이미 추출된 초기표본에서 다시 이중추출된 각 표본쌍은 단순임의추출로 뽑혔으며 고비용을 들여 추출되어 초기표본보다는 매우 정확하게 분류되었다고 가정한다. 예를 들면 전화조사나 우편조사 혹은 인터넷을 이용한 조사 등의 저비용의 대규모 여론조사의 경우가 초기표본이며 이렇게 추출된 초기표본중 이중추출된 부표본은 소규모로 단위표본당 고비용을 들여 일대일 면접법(face-to-face interview)으로 조사되어 매우 정확히 분류된 자료인 경우이다. N개의 초기표본과 n개의 부표본의 표본공간(sample space)을 각각 $\mathcal{R}$과 $\mathcal{X}$라 표기하면 두 표본공간은 $\mathcal{X} \subset \mathcal{R}$인 관계가 있다. 따라서 초기표본에서 부표본을 제외한 $N^* = N - n$개의 표본을 차표본(different sample)이라 정의하고 차표본의 표본공간을 $\mathcal{R}^* = \mathcal{R} - \mathcal{X}$로 표기한다. 초기표본이 확률표본(random sample)이기 때문에 각각의 표본은 독립이며, 부표본과 차표본도 독립적인 표본이 된다.

본 논문에서는 초기표본으로 이루어진 분할표를 원분할표(original table), 이중추출에 의해 추출된 부표본에 대한 분할표를 완비분할표(complete table), 그리고 원분할표에서 완

비분할표에 대응되는 오분류 가능성이 있는 변수에 대한 주변합을 제외한 차표본에 대한 분할표를 불확실 분할표(uncertain table)로 정의한다. 또한 불확실 분할표의 각 칸의 도수는 원분할표에서 오분류 가능성이 없는 변수 X의 주변합에 대한 조건부확률에 불확실 분할표의 주변합인 N^*_{i+}을 곱하여 다음과 같이 정의하며, 변수 X에 대한 원분할표에서의 조건부확률과 불확실 분할표에서의 조건부확률이 동일함을 알 수 있다.

$$N^*_{ij} = N^*_{i+}\frac{N_{ij}}{N_{i+}}.$$

표 2.1은 위의 개념을 소개하고자 2×2분할표인 경우를 설명하고 있다. N개의 초기표본 자료를 정리한 원분할표에서 범주형 변수 Y만이 오분류 가능성이 있는 경우이다. 초기표본에서 이중추출되었고 정확하게 분류된 부표본 자료 n개를 완비분할표에서 정리하였다. 완비분할표에서의 모든 칸들의 값 n_{ij}는 정확하다고 가정하고 원분할표에서는 X의 주변합만이 정확하기 때문에, 초기표본에서 부표본을 제외한 차표본 자료를 정리한 불확실 분할표에서도 X의 주변합 N^*_{i+}만을 정확한 정보로 간주한다. 따라서 불확실 분할표의 칸값 N^*_{ij}는 오분류 가능성이 있는 값이므로 표 2.1의 오른쪽 표인 추정분할표에서 ?으로 표현된 각 칸비율을 추정하여 불확실 분할표에서 대응하는 칸비율과 비교하여 오분류 여부를 파악하고자 한다.

표 2.1은 Little과 Rubin(1987, p237)에서 무응답이나 결측값이 존재하는 자료형태와 유사하다. 따라서 오분류의 문제를 해결하기 위해서 Little과 Rubin(1987)이 제안한 결측값(또는 무응답)을 추정하는 방법을 이용하여 접근할 수 있다. 단위표본당 조사소요비용, 외부의 정보 등과 더불어 완비분할표의 정보와 불확실 분할표에서 정확히 분류된 X의 주변합의 정보를 핵심적으로 사용하여 추정분할표에서 칸값 또는 칸비율을 추정하여 오분류의 정도를 파악하고자 한다.

표 2.1: 분할표의 형태

		원분할표 ($\mathcal{R}$)			완비분할표 ($\mathcal{X}$)			불확실 분할표 ($\mathcal{R}^* = \mathcal{R} - \mathcal{X}$)			추정분할표		
		Y			Y			Y			Y		
		j=1	j=2	주변합	j=1	j=2	주변합	j=1	j=2	주변합	j=1	j=2	주변합
X	i=1	N_{11}	N_{12}	N_{1+}	n_{11}	n_{12}	n_{1+}	N^*_{11}	N^*_{12}	N^*_{1+}	?	?	N_{1+}
	i=2	N_{21}	N_{22}	N_{2+}	n_{21}	n_{22}	n_{2+}	N^*_{21}	N^*_{22}	N^*_{2+}	?	?	N_{2+}
주변합		N_{+1}	N_{+2}	N	n_{+1}	n_{+2}	n	N^*_{+1}	N^*_{+2}	N^*	?	?	N

3. 베이지안 추정

표 2.1의 분할표를 $I \times J$분할표로 확장하여 각 분할표에 대한 표본도수 공간을 고려하여 보자. 우선 초기표본에서 고비용을 들여 이중추출된 부표본에 대한 완비분할표에서 각

칸값들의 집합인 완비분할표의 표본도수 공간을 다음과 같이 정의한다.

$$S_c = \{(i,j) = n_{ij} \in \mathcal{X},\ i = 1,\ldots,I,\ j = 1,\ldots,J\},$$

여기서 S_c의 아래첨자 c는 완비(complete)의 약자이다. 오분류 가능성이 없는 변수 X의 주변합만이 정확한 정보를 갖는 불확실 분할표에서 각 칸값들의 집합인 불확실 분할표의 표본도수 공간을 다음과 같이 정의한다.

$$S_{\mathcal{R}^*} = \{(i,j) \quad = N^*_{ij} \in \mathcal{R}^*,\ i = 1,\ldots,I,\ j = 1,\ldots,J\}.$$

완비분할표와 불확실 분할표의 합의 형태로 나타나는 초기표본에 대한 원분할표의 표본도수 공간을 다음과 같이 나타낼 수가 있다.

$$S_{\mathcal{R}} = S_c + S_{\mathcal{R}^*} = \{(i,j) \quad = N_{ij} \in \mathcal{R},\ i = 1,\ldots,I, j = 1,\ldots,J\}.$$

완비분할표의 표본도수 공간인 S_R은 완비분할표의 표본도수 공간 S_c와 불확실 분할표의 표본도수 공간 S^*_R로 분할하여 나타낼 수 있다. 즉 표 2.1에서 오분류 가능성이 없는 변수 X에 대한 주변합은 $N_{i+} - n_{i+} = N^*_{i+}$인 관계를 갖으나 칸 도수는 동일한 성질을 갖지 않는다 (즉 $N_{ij} - n_{ij}$는 항상 N^*_{ij}가 아니다). 왜냐하면 초기표본의 표본공간은 부표본의 표본공간을 포함하지만 부표본의 도수는 새롭게 정확하게 조사된 자료이기 때문에 초기표본에 대응되는 부표본의 값이 동일하게 나타나지 않기 때문이다. 따라서 원분할표의 표본도수 공간과 불확실 분할표의 표본도수 공간은 종속적인 관계를 갖으나, 원분할표의 표본도수 공간과 완비분할표의 표본도수 공간은 독립이 된다.

무응답값의 추정에 결측값 메카니즘을 이용하기 위하여 추정과정에서 고려되는 표본도수 공간을 다음과 같이 완비분할표와 불확실 분할표의 표본도수 공간의 합의 형태로 정의하자.

$$S_{est} = S_c + \omega \cdot S_{R^*},$$

여기서 아래 첨자 est는 추정(estimation)의 약자이며 추정을 위해 사용되는 표본도수 공간을 나타낸다. ω은 지시함수(indicator function)로 만약 1의 값을 갖으면 완비분할표와 불확실 분할표를 모두 고려한 표본공간인 $S_{est} = S_c + S_{R^*} = S_R$ 으로 원분할표의 표본도수 공간과 동일하게 된다. 반면 ω의 값이 0의 값을 갖으면 추정을 위한 표본도수 공간 S_{est}은 완비분할표의 표본공간인 S_c이 된다.

우선 $\omega = 0$이 되어 무시할 수 있는 무응답(IN)의 경우를 살펴보면 추정과정에 완비분할표의 표본도수 공간만을 고려하며, 완비분할표에서의 각 칸의 도수는 다음과 같이 다항분포를 따른다고 가정하자.

$$P[S_c|\underline{\theta}] = \prod_i^I \prod_j^J \frac{n!}{n_{ij}!} {\theta_{ij}}^{n_{ij}},$$

여기서 $n = \sum_{i,j} n_{ij}$이며 모수 $\theta_{ij} = P[X = i, Y = j] > 0$는 분할표에 칸의 확률로 벡터형태는 $\underline{\theta} = (\theta_{11}, \ldots, \theta_{IJ}) = (\theta_{ij})$, $\sum_{i,j} \theta_{ij} = 1$으로 나타낸다. 이때 $\underline{\theta} = (\theta_{ij})$의 사전분포를 다항분포의 공액분포족인 Dirichlet분포로 가정하는데 우선 θ_{ij}는 1차원의 Dirichlet분포인 Beta분포를 따르며 θ_{ij}의 확률밀도함수는 다음과 같고, $\theta_{ij} \sim D(\alpha_{ij}, \alpha - \alpha_{ij})$로 표기된다.

$$P[\theta_{ij}] = \frac{\Gamma(\alpha)}{\Gamma(\alpha_{ij})\Gamma(\alpha - \alpha_{ij})} {\theta_{ij}}^{\alpha_{ij}-1}(1-\theta_{ij})^{\alpha-\alpha_{ij}-1},$$

여기서 α_{ij}는 Dirichlet분포의 모수이며 $\alpha_{ij} \geq 0$, $\alpha > \alpha_{ij}$ 이다. 그리고 모수벡터 $\underline{\theta}$에 대한 사전분포는 다음과 같이 $I \times J$차의 Dirichlet분포로 가정한다.

$$P[\underline{\theta}] = \prod_{i}^{I} \prod_{j}^{J} \frac{\Gamma(\alpha)}{\Gamma(\alpha_{ij})} {\theta_{ij}}^{\alpha_{ij}-1},$$

여기서 $\sum_{i,j} \alpha_{ij} = \alpha$이다. 또한 θ_{ij}의 사후분포는 모수벡터의 사전분포와 마찬가지로 다음과 같은 Dirichlet분포를 따르며 $\theta_{ij}|S_c \sim D(\alpha_{ij} + n_{ij}, \alpha + n - \alpha_{ij} - n_{ij})$ 으로 표기된다.

$$P[\theta_{ij}|S_c] = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha_{ij} + n_{ij})\Gamma(\alpha + n - \alpha_{ij} - n_{ij})} {\theta_{ij}}^{\alpha_{ij}+n_{ij}-1}(1-\theta_{ij})^{\alpha+n-\alpha_{ij}-n_{ij}-1}.$$

본 논문에서는 $I \times J$의 분할표에서 Y에만 오분류의 가능성을 가정하고 있기 때문에 정확하게 분류된 X의 i번째 범주의 주변합에 대한 Y의 j번째 범주에 대한 조건부확률에 관심을 갖는다. 조건부확률 $\theta_{j|i}$의 사전분포와 사후분포를 고려해 보자. 주변합에 대한 사전분포는 $\theta_{i+} \sim D(\alpha_{i+}, \alpha - \alpha_{i+})$와 $\theta_{+j} \sim D(\alpha_{+j}, \alpha - \alpha_{+j})$이 되며, 여기서 $\theta_{i+} = \sum_j \theta_{ij}$, $\theta_{+j} = \sum_i \theta_{ij}$, $\alpha_{i+} = \sum_j \alpha_{ij}$, $\alpha_{+j} = \sum_i \alpha_{ij}$이며 θ_{i+}와 $\theta_{j|i}$, θ_{+j}와 $\theta_{i|j}$는 상호독립이다. 따라서 i번째 행범주의 주변합에 대한 j번째 응답칸의 확률은 $\theta_{j|i} = \theta_{ij}/\theta_{i+}$이며, 반대로 j번째 행범주의 주변합에 대한 i번째 응답칸의 확률은 $\theta_{i|j} = \theta_{ij}/\theta_{+j}$이다. 또 $\theta_{j|i}$와 $\theta_{i|j}$의 사전분포도 $D(\alpha_{ij}, \alpha_{i+} - \alpha_{ij})$와 $D(\alpha_{ij}, \alpha_{+j} - \alpha_{ij})$인 Dirichlet분포를 갖게 된다. 따라서 무시할 수 있는 무응답(IN)의 경우에 $\theta_{j|i}$에 대한 사후분포는 완비분할표만을 고려하여 다음과 같은 분포를 갖는다(Sebastiani와 Ramoni 1997).

$$\theta_{j|i}|S_c \sim D(\alpha_{ij} + n_{ij}, \alpha_{i+} + n_{i+} - \alpha_{ij} - n_{ij}).$$

다음은 무응답이 무시할 수 없는 무응답(NIN)이라 고려되는 경우를 고려하여 보자. 즉 $S_{est} = S_c + \omega \cdot S_{R^*}$에서 $\omega = 1$인 경우로 $S_{est} = S_R$이 되어 완비분할표와 불확실 분할표의 정보를 이용하게 된다. 원분할표에서의 각 칸의 도수는 무시할 수 있는 무응답의 경우와 동일하게 다음과 같은 다항분포를 따른다고 가정하자.

$$P[S_{\mathcal{R}}|\underline{\theta}] = \prod_{i}^{I} \prod_{j}^{J} \frac{N!}{N_{ij}!} {\theta_{ij}}^{N_{ij}},$$

여기서 $N = \sum_{i,j} N_{ij}$이다. 이때 θ_{ij}의 사전분포를 다항분포의 공액분포족인 Dirichlet분포로 고려하면 $\underline{\theta}$에 대한 사전분포도 역시 $I \times J$차의 Dirichlet분포를 따르게 된다. 원분할표

에서 θ_{ij}의 사후분포를 고려하면 다음과 같다.

$$\theta_{ij}|S_{\mathcal{R}} \sim D(\alpha_{ij}+N_{ij},\ \alpha+N-\alpha_{ij}-N_{ij}).$$

변수 X의 i번째 범주의 주변합에 대한 변수Y의 j번째 범주의 조건부확률 $\theta_{j|i}$의 사후분포를 구하여 보면 Dirichlet분포를 따르게 된다.

$$\theta_{j|i}|S_{\mathcal{R}} \quad \sim \quad D(\alpha_{ij}+N_{ij},\ \alpha_{i+}+N_{i+}-\alpha_{ij}-N_{ij}). \tag{3.1}$$

위의 사후분포식에서 초기표본에 대한 원분할표에서 조건부확률 $\theta_{j|i}$는 완비분할표의 표본과 불확실 분할표의 표본에 대한 사후분포로도 유도할 수 있다. 원분할표에서 사후조건부확률 $\theta_{j|i}$에 대한 평균은 다음과 같다.

$$E(\theta_{j|i}|S_{\mathcal{R}}) \quad = \quad \frac{\alpha_{ij}+N_{ij}}{\alpha_{i+}+N_{i+}}. \tag{3.2}$$

원분할표에서 $\theta_{j|i}$의 기대값인 (3.2)식에서 각 칸의 도수 N_{ij}는 $0 \le N_{ij}-n_{ij} \le N^*_{i+}$에 의하여 Sebastiani와 Ramoni(1997)는 사후 조건부확률 $\theta_{j|i}|S_{\mathcal{R}}$에 대한 추정값을 다음과 같은 구간으로 나타내었다.

$$\hat{\theta_{j|i}}^{L} = \frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}+N^*_{i+}} \le \theta_{j|i}|S_{\mathcal{R}} \le \frac{\alpha_{ij}+n_{ij}+N^*_{i+}}{\alpha_{i+}+n_{i+}+N^*_{i+}} = \hat{\theta_{j|i}}^{U}.$$

그리고 $\theta_{j|i}|S_{\mathcal{R}}$의 점추정값은 상한값인 $\hat{\theta_{j|i}}^{U}$와 하한값인 $\hat{\theta_{j|i}}^{L}$의 선형결합(linear combination)에 의해서 구하였다.

$$\begin{aligned}
\hat{\theta}_{j|i}|S_{\mathcal{R}} &= \phi_{j|i}\hat{\theta_{j|i}}^{U} + (1-\phi_{j|i})\hat{\theta_{j|i}}^{L} \\
&= \frac{\alpha_{ij}+n_{ij}+\phi_{j|i}N^*_{i+}}{\alpha_{i+}+n_{i+}+N^*_{i+}} \qquad (3.3) \\
&= \frac{\alpha_{i+}+n_{i+}}{\alpha_{i+}+n_{i+}+N^*_{i+}}\frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} + \frac{N^*_{i+}}{\alpha_{i+}+n_{i+}+N^*_{i+}}\phi_{j|i}\ , \qquad (3.4)
\end{aligned}$$

여기서 외부정보(external information)인 $\phi_{j|i}$는 $\sum_j \phi_{j|i}=1$이며 결측값 패턴에 대한 정보이다. 위와 같은 구간추정량과 점추정량을 Sebastiani와 Ramoni(1997)는 Bound와 Collapse라고 하였다. 원분할표에서 사후추정량 $\hat{\theta}_{j|i}$는 완비분할표로부터 얻어진 $\theta_{j|i}$의 최고가능도 추정량(MLE)인 $(\alpha_{ij}+n_{ij})/(\alpha_{i+}+n_{i+})$와 외부정보 $\phi_{j|i}$의 가중평균형태로 나타난다(Little과 Rubin 1987). N^*_{i+}이 감소한다면 완비분할표로부터 얻어진 $\theta_{j|i}$의 추정값 $(\alpha_{ij}+n_{ij})/(\alpha_{i+}+n_{i+})$은 $\phi_{j|i}$보다 높이 가중되며, $N^*_{i+}=0$이라면 원분할표에서의 사후추정량은 $\hat{\theta}_{j|i}=(\alpha_{ij}+n_{ij})/(\alpha_{i+}+n_{i+})$이 되며 이는 완비분할표에서의 사후추정량 $\theta_{j|i}$의 평균 $E(\theta_{j|i}|S_c)$이 된다. 반면에 N^*_{i+}이 증가한다면 사후추정량 $\hat{\theta}_{j|i}$은 외부정보 $\phi_{j|i}$에 근접한다는 것을 파악할 수 있다. 따라서 원분할표에서 사후조건부확률 $\theta_{j|i}$의 평균인 (3.2)식에서 (3.3)식으로 변경될 수 있기 때문에 $\theta_{j|i}$의 사후분포를 $\phi_{j|i}$에 의해서 다시 표현하면 (3.1)식에서 N_{ij}가 $n_{ij}+\phi_{j|i}N^*_{i+}$로 변경되어 다음과 같이 유도할 수 있다.

$$\theta_{j|i}|S_{\mathcal{R}} \sim D(\alpha_{ij}+n_{ij}+\phi_{j|i}N^*_{i+},\ \alpha_{i+}+n_{i+}+N^*_{i+}-\alpha_{ij}-n_{ij}-\phi_{j|i}N^*_{i+}).$$

이와 같은 상황에서는 원분할표에서 추가적인 외부정보 $\phi_{j|i}$에 의해서 새롭게 표현된 $\theta_{j|i}$는 $P[Y=j|X=i,\underline{\alpha},\phi_{j|i},S_{\mathcal{R}}]$으로 정의되며 이때 $\theta_{j|i}$의 사후분포의 평균과 분산인 $E(\theta_{j|i}|S_{\mathcal{R}})$와 $V(\theta_{j|i}|S_{\mathcal{R}})$는 다음과 같이 구한다.

$$\begin{aligned} E(\theta_{j|i}|S_{\mathcal{R}}) &= \frac{\alpha_{ij}+n_{ij}+\phi_{j|i}N^*_{i+}}{\alpha_{i+}+n_{i+}+N^*_{i+}} \\ &\equiv \hat{\theta}_{j|i}, \end{aligned} \tag{3.5}$$

$$\begin{aligned} V(\theta_{j|i}|S_{\mathcal{R}}) &= \frac{E(\theta_{j|i}|S_{\mathcal{R}})(1-E(\theta_{j|i}|S_{\mathcal{R}}))}{\alpha_{i+}+n_{i+}+N^*_{i+}+1} \\ &= \frac{\hat{\theta}_{j|i}(1-\hat{\theta}_{j|i})}{\alpha_{i+}+n_{i+}+N^*_{i+}+1} \\ &\equiv \hat{V}(\hat{\theta}_{j|i}). \end{aligned} \tag{3.6}$$

따라서 무시할 수 있는 무응답(IN)와 무시할 수 없는 무응답(NIN)의 경우를 모두 고려한 일반적인 $\theta_{j|i}$의 추정량은 다음과 같으며,

$$\begin{aligned} \hat{\theta}_{j|i} &= E(\theta_{j|i}|S_{est}) \\ &= \frac{\alpha_{ij}+n_{ij}+\omega\cdot\phi_{j|i}N^*_{i+}}{\alpha_{i+}+n_{i+}+\omega\cdot N^*_{i+}}, \end{aligned} \tag{3.7}$$

추정분산 $\hat{V}(\hat{\theta}_{j|i})$도 기대베이지안 추정량(expected bayesian estimator)인 $\hat{\theta}_{j|i}$의 함수로 나타난다. 또한 이렇게 추정된 기대베이지안 추정량 $\hat{\theta}_{j|i}$은 외부정보 $\phi_{j|i}$와 사전모수벡터 $\underline{\alpha}$의 함수형태이다.

4. 정보의 종류

4.1. 외부정보

외부정보는 결측값 매카니즘에 대한 정보로 불확실 분할표에서 오분류 가능성이 없는 변수 X에 대한 주변합이 각 칸에 배분될 확률로 다음과 같이 나타난다.

$$\phi_{j|i} = P[Y=j|X=i,Y=?,\underline{\alpha}],$$

여기서 $\sum_j \phi_{j|i}=1$이며, 이 경우에도 무시할 수 있는 무응답과 무시할 수 없는 무응답의 경우에 따라 외부정보가 구분되는데, 무시할 수 있는 무응답이라면 외부정보는 다음과 같이 완비분할표의 정보와 사전모수에 의해 표현될 수 있다.

$$\phi_{j|i} = \frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}}. \tag{4.1}$$

정리 4.1 무시할 수 있는 무응답인 경우에 (4.1)에서 정의된 외부정보 $\phi_{j|i}$으로 설명되는 $\hat{\theta}_{j|i}|S_R$은 기대베이지안 추정량 $\hat{\theta}_{j|i}|S_c$은 동일하다.

증명: $\theta_{j|i}$의 추정식 (3.4)에서 $\phi_{j|i}$에 (4.1)식을 대입하면,

$$\begin{aligned}\hat{\theta}_{j|i} &= \frac{\alpha_{i+}+n_{i+}}{\alpha_{i+}+N_{i+}}\frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} + \frac{N^*_{i+}}{\alpha_{i+}+N_{i+}}\phi_{j|i} \\ &= \frac{\alpha_{i+}+n_{i+}}{\alpha_{i+}+N_{i+}}\frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} + \frac{N^*_{i+}}{\alpha_{i+}+N_{i+}}\frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} \\ &= \frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} = E(\theta_{j|i}|S_c).\end{aligned}$$

그러므로 무시할 수 있는 무응답의 경우에 기대베이지안 추정량 $\hat{\theta}_{j|i}$은 $E(\theta_{j|i}|S_c)$ 으로 구할 수 있다. □

반면 불확실 분할표의 주변합을 무시할 수 없는 무응답이라 고려하면, 외부정보는 초기 표본조사 비용과 이중추출시의 조사비용을 고려하여 사용한다. 즉 불확실 분할표의 자료 한개를 조사할 때 소요되는 비용 c_2와 완비분할표의 자료 한개를 조사할 때의 비용을 c_1이라 하고($c_1 > c_2$), $C = c_1 + c_2$라 하면 다음과 같은 외부정보를 고려할 수 있다.

$$\phi_{j|i} = \frac{c_1}{C}\frac{\alpha_{ij}+n_{ij}}{\alpha_{i+}+n_{i+}} + \frac{c_2}{C}\frac{\alpha_{ij}+N^*_{ij}}{\alpha_{i+}+N^*_{i+}} \tag{4.2}$$

4.2. 사전모수

Good(1968)은 $I \times J$분할표에 대한 사전분포의 모수인 사전모수를 $\alpha_{ij} = (I \times J)^{-1}$인 균일분포로 제안하였다. 균일한 사전모수를 사용할 경우에 추정에 편의가 발생할 수 있으며 초기표본이나 이중추출된 표본의 크기가 커질 경우 사전모수는 기대베이지안 추정량 $\hat{\theta}_{j|i}$에 거의 영향을 주지 못한다는 단점이 있다. 따라서 본 논문에서는 사전모수를 완비분할표와 불확실 분할표에서의 정보를 이용한 경험적 베이지안 사전모수(empirical bayesian prior parameter)로 사용하였다.

우선 무시할 수 있는 무응답(IN)의 경우는 다음과 같은 사전모수를 고려할 수 있다.

표 4.1: IN인 경우에 대한 사전모수

$n/(I \times J)$	n_{ij}	$\hat{n}_{ij}$

표 4.1의 사전모수중 첫번째 $n/(I \times J)$은 무작위 사전모수(improper prior)로 완비분할표의 도수가 균일분포를 따른다는 가정하에서의 사전모수이며, n_{ij}는 완비분할표의 자체의 도수 즉, 포화모형(saturated model)의 가정하에서의 추정도수이며, 그리고 $\hat{n}_{ij}$은 독립성 모형(independent model)의 가정하에 추정된 도수이다. 사전모수에 관한 사전정보(pre-information)가 있다면 사전모수의 분포를 설정하는데 이용을 할 수 있으나, 사전정보가

없을 때에는 완비분할표의 표본에서 얻은 정보를 이용하여 경험적 사전모수로 이용한다. 무시할 수 있는 무응답의 경우에는 완비분할표의 정보만을 고려하며, 완비분할표는 정확하게 분류되었다고 가정하기 때문에 표 4.1에 제시된 사전모수중 완비분할표의 정보를 가장 잘 나타내는 사전모수는 n_{ij}이다. 사전모수를 $\alpha_{ij} = n_{ij}$로 설정하면, 외부정보와 베이지안 추정량 모두 $\hat{\theta}_{j|i} = \phi_{j|i} = n_{ij}/n_{i+}$이 되어 완비분할표만을 고려한 경우가 된다.

표 4.2: NIN인 경우에 대한 사전모수

nN^*_{ij}/N^*	$n[N_{ij}/N]$	$(n/2)[n_{ij}/n + N^*_{ij}/N^*]$	$n\left(\frac{c_1}{C}\frac{n_{ij}}{n} + \frac{c_2}{C}\frac{N^*_{ij}}{N^*}\right)$

무시할 수 없는 무응답(NIN)의 경우에 대한 사전모수는 표 4.2과 같이 고려할 수 있다. 이 경우에도 n_{ij}와 N^*_{ij} 대신에 독립성모형하에서의 추정량 $\hat{n}_{ij}$와 $\hat{N}^*_{ij}$를 이용한 사전모수를 고려할 수도 있으나, 무시할 수 있는 무응답의 경우와 동일하게 포화모형인 경우에 완비분할표와 불확실 분할표자체의 정보를 가장 잘 반영한다고 할 수 있다. 따라서 독립성모형하의 표 4.2에서는 $\hat{n}_{ij}$와 $\hat{N}^*_{ij}$를 고려하지 않고 포화모형인 경우에 고려할 수 있는 사전모수만을 제시하였다. 표 4.2에서 첫번째 제시된 사전모수는 불확실 분할표만의 정보를 이용한 것이며 두번째는 원분할표의 정보를 고려한 것으로 완비분할표에 대한 정보가 손실된다. 세번째는 불확실 분할표의 정보와 완비분할표의 정보에 대한 평균형태로 정보의 정확도에 따른 가중을 고려하지 못했다. 따라서 마지막으로 제시되는 사전모수는 완비분할표의 정보와 불확실 분할표의 정보를 각 단위당 조사비용에 의해 가중을 주었으며 위의 네 가지 사전모수중 가장 바람직한 경험적 사전모수라 고려된다. 또한 마지막 경우의 사전모수는 다음과 같이 추가적인 정보를 이용하여 더욱 현실적인 사전모수로 표현할 수 있다.

$$\alpha_{ij} = n\left[k\left(\frac{c_1}{C}\frac{n_{ij}}{n} + \frac{c_2}{C}\frac{N^*_{ij}}{N^*}\right) + (1-k)\alpha^*_{ij}\right], \tag{4.3}$$

여기서 사전정보(pre-information) α^*_{ij}는 다른 연구기관에서 실시한 유사한 조사의 결과나 혹은 동일한 조건하에서 전년도에 조사된 결과 등으로 확률형태를 갖는다. 사전정보의 값은 신뢰정도에 따라 분석전에 임의로 설정할 수 있으며 사전 설정된 가중값 $k \in [0,1]$를 사용하여 사전모수 (4.3)식을 구할 수 있다. 만약 사전정보가 전혀 없다면 $k = 1$로 하여 오직 경험적인 사전모수만을 사용할 수 있을 것이다.

5. MC-통계량

$I \times J$분할표의 오분류 여부를 검정하기 위해서 오분류 되었다고 간주되는 분할표를 이중추출법에 의해 수집된 정보와 단위표본당 소요비용, 외부정보, 사전정보 등을 이용하여 추정분할표의 칸에 대응하는 베이지안 추정량인 조건부확률 $\theta_{j|i} = P[Y = j|X = i,\ \underline{\alpha},\ \phi_{j|i},\ S_{est}]$을 추정하였다. 또 원분할표에서 완비분할표의 칸값을 제외한 불확실 분할표의 칸값에 대응하는 조건부확률 $p_{j|iN^*} = P[Y = j|X = i, S_{\mathcal{R}^*}]$을 구하여 베이지안 추정

량과 동일하다는 다음의 가설을 검정하고자 한다.

$$H_0 : p_{j|iN^*} = \theta_{j|i}, \quad H_1 : p_{j|iN^*} \neq \theta_{j|i}. \qquad i = 1, \ldots, I, \ j = 1, \ldots, J. \tag{5.1}$$

여기서 $p_{j|iN^*}$의 추정량의 경우에는 불확실 분할표의 표본도수 공간인 S_{R^*}을 사용하여 추정하였으며 칸도수는 원분할표의 조건부확률을 이용하였기 때문에 언제나 $\hat{p}_{j|iN^*} = N_{ij}/N_{i+}$으로 나타나며 $\theta_{j|i}$의 추정량의 하나로서 $\hat{p}_{j|iN^*}$을 고려하지 않는다.

그러므로 가설 (5.1)에서 $p_{j|iN^*} = \theta_{j|i}$을 귀무가설로 설정하여도 기대베이지안 추정값 $\theta_{j|i}$와 불확실분할표의 추정조건부확률 $p_{j|iN^*}$와 동일한 경우는 나타나지 않는다. 만약 동일한 값을 갖는다면 표본도수 공간이 S_R인 원분할표의 조건부확률과 표본도수 공간이 S_c인 완비분할표의 조건부확률값이 완전히 일치하여 오분류가 존재하지 않는 경우임을 인식할 수 있다.

여기서는 위의 가설 (5.1)을 검정하기 위해 χ^2-분포를 따르는 대안적인 검정통계량을 제안한다. 가설 (5.1)를 검정하기 위하여 제안하는 MC-통계량(Misclassification statistic)은 표본집합 $S_{\mathcal{R}}$하에서 추정한 $\hat{\theta}_{j|i}$와 표본집합 $S_{\mathcal{R}^*}$하에서 추정한 $\hat{p}_{j|iN^*}$의 비율차이에 대하여 검정하여 표본공간이 달라지는데 따른 문제점을 제거하였다. 그러므로 MC-통계량은 일종의 비율차 검정통계량으로 $I \times J$인 분할표에 대해서 다음과 같이 유도할 수 있다.

정리 5.1 원분할표에 대하여 베이지안 추정법을 이용한 조건부확률 $\theta_{j|i} = P[Y = j|X = i, \underline{\alpha}, \phi_{j|i}, S_{est}]$와 불확실 분할표에 대한 조건부확률 $p_{j|iN^*} = P[Y = j|X = i, S_{\mathcal{R}^*}]$ 의 동일함의 여부에 따라 오분류된 자료임을 파악하기 위한 가설 (5.1)에 대한 검정통계량은 다음과 같이 정의되며, MC-통계량은 자유도가 $I \times (J-1)$인 χ^2분포를 따른다.

$$MC(\theta_{j|i}) = \sum_{i}^{I} \sum_{j}^{J} \frac{(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})^2}{\hat{V}(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})}, \qquad i = 1, \ldots, I, \ j = 1, \ldots, J, \tag{5.2}$$

여기서 $\hat{p}_{j|iN^*} = N^*_{ij}/N^*_{i+}$와 $\hat{\theta}_{j|i} = (\alpha_{ij} + n_{ij} + \phi_{j|i}N^*_{i+})/(\alpha_{i+} + n_{i+} + N^*_{i+})$이며

$$\hat{V}(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i}) = \frac{\hat{p}_{j|iN^*}(1 - \hat{p}_{j|iN^*})}{N^*_{i+}} + \frac{\hat{\theta}_{j|i}(1 - \hat{\theta}_{j|i})}{\alpha_{i+} + n_{i+} + N^*_{i+} + 1}.$$

증명: $\hat{p}_{j|iN^*}$은 불확실 분할표에서 오분류가 존재하지 않는 변수의 i번째 범주에 대한 조건부 칸확률의 추정값 N^*_{ij}/N^*_{i+}이며 이때 불확실 분할표의 칸 도수는 $N^*_{ij} = (N_{ij}/N_{i+}) \cdot N^*_{i+}$에 의해서 다시 N_{ij}/N_{i+}으로 구할 수 있다. 또한 추정량 $\hat{p}_{j|iN^*}$의 확률변수는 N^*_{ij}으로 모수가 N^*_{i+}와 $p_{j|iN^*}$인 이항분포를 따르므로 $\hat{p}_{j|iN^*}$의 분산은 $\hat{p}_{j|iN^*}(1 - \hat{p}_{j|iN^*})/N^*_{i+}$이다.

반면 $\hat{\theta}_{j|i}$ 은 $\theta_{j|i}$의 기대베이지안 추정량으로 (3.5)식이며 추정분산은 (3.6)식으로부터 얻을 수 있다. $\hat{\theta}_{j|i} = (\alpha_{ij} + n_{ij} + \phi_{j|i}N^*_{i+})(\alpha_{i+} + n_{i+} + N^*_{i+})$의 확률변수는 사전모수 α_{ij}와 외부정보 $\phi_{j|i}$ 그리고 주변합들이 주어진 조건 하에서 n_{ij}이 된다. 따라서 각 추정량 $\hat{p}_{j|iN^*}$와 $\hat{\theta}_{j|i}$의 확률변수는 각각 N^*_{ij}와 n_{ij}이되는데 이는 3절에서 언급과 동일하게 오분류 가능성이 없는 변수 X에 대한 주변합은 $N_{i+} - n_{i+} = N^*_{i+}$인 관계를 갖지만 각 분할표의 칸의 도수는 동일한 성질을 갖지 않는다. 왜냐하면 초기표본의 표본공간은 부표본의 표본공간을 포

함하지만 부표본의 도수는 새롭게 조사된 자료이기 때문에 초기표본에 대응되는 부표본의 값이 동일하게 나타나지 않기 때문이다. 즉, 원분할표의 표본도수 공간과 불확실 분할표의 표본도수 공간은 종속적인 관계를 갖으며, 원분할표의 표본도수 공간과 완비분할표의 표본도수 공간은 독립이 된다. 따라서 N_{ij}^*와 n_{ij}이 독립이 되므로 $\hat{p}_{j|iN^*}$와 $\hat{\theta}_{j|i}$은 서로 독립이다. 그러므로 추정분산 $\hat{V}(\hat{p}_{j|i'N^*} - \hat{\theta}_{j|i})$을 위와 같이 구할 수 있다. 가설 (5.1)에 대한 귀무가설하에서의 통계량

$$Z = \frac{(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})}{\sqrt{\hat{V}(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})}}$$

은 근사적으로 평균이 0이고 분산이 1인 표준정규분포를 따르므로 MC-통계량은 $\sum_{j=1}^{J} p_{j|iN^*} = \sum_{j=1}^{J} \theta_{j|i} = 1$인 조건에 의하여 자유도가 $I \times (J-1)$인 χ^2-분포를 따르게 된다. □

만약 오분류되어 있다고 고려되는 불확실 분할표에서 조건부 칸확률과 추정된 분할표에서의 기대베이지안 추정값이 동일하다는 귀무가설을 채택하게 되면 오분류가 없는 것으로 판단되며, 귀무가설을 기각하게 되면 불확실 분할표에는 오분류가 존재하며 따라서 원분할표가 오분류된 것으로 고려된다. 오분류 되었다고 판단되는 경우에는 원분할표의 칸값을 추정된 기대베이지안 값을 이용하여 대체하여 사용할 수 있다.

보조정리 5.1 불확실 분할표에서 특정한 $i(=1,\ldots,I)$번째 범주에 대하여만 오분류 여부를 검정하기 위한 가설

$$H_0 : p_{j|iN^*} = \theta_{j|i}, \quad H_1 : p_{j|iN^*} \neq \theta_{j|i}. \qquad j = 1,\ldots,J, \tag{5.3}$$

에 대한 검정통계량은 다음과 같다.

$$MC_i(\theta_{j|i}) = \sum_{j}^{J} \frac{(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})^2}{\hat{V}(\hat{p}_{j|iN^*} - \hat{\theta}_{j|i})}. \tag{5.4}$$

위에서 정의한 MC_i-통계량은 $\sum_{j=1}^{J} p_{j|iN^*} = \sum_{j=1}^{J} \theta_{j|i} = 1$인 조건에 의해 자유도가 $J-1$인 χ^2-분포를 따른다.

정리 5.1에서 제안된 MC-통계량은 보조정리 5.1의 MC_i-통계량의 합으로 나타난다. 그러므로 MC-통계량은 오분류 가능성이 없는 행변수의 모든 범주에 대한 오분류 여부를 검정할 수 있으며, MC_i-통계량은 하나의 행변수 범주에 대하여 검정할 수 있다. 다음 절에서 모의실험을 통하여 제안된 통계량들의 성질을 연구해 보자.

6. 모의실험

모의실험은 무시할 수 있는 무응답(IN)과 무시할 수 없는 무응답(NIN)의 경우로 나누어 실시되었다. $I = J = 2$인 가장 간단한 2×2분할표에서 원분할표의 표본크기를 1,000개

부터 10,000개까지 변화시켰으며 이중추출율을 1%~30%까지 변화시면서 자료를 생성하였다. 그리고 원분할표에서 변수 X의 각 범주에 대한 주변합을 0.1부터 0.9까지 변화시켰으며, 변수 X의 각 범주에서 Y의 첫번째 범주에 대한 조건부확률을 변화시키면서 실험하였다. 오분류를 가정하기 위하여 원분할표의 조건부확률에 대응되는 완비분할표의 조건부확률, 즉 변수 X의 첫번째 범주와 두번째 범주를 조건으로 하는 Y에 첫번째 범주의 조건부확률에 대하여 각각 D_1과 D_2만큼의 오분류율을 설정하여 원분할표와 완비분할표의 각 칸 값들을 생성하였다. 오분류율 $D_i(i = 1, 2)$는 1%~3%까지 변화시키면서 MC-통계량의 변화를 살펴보았다.

6.1. 표본크기변화에 따른 MC-통계량

원분할표의 표본크기를 1,000개부터 10,000개까지 변화시키면서 MC-통계량의 변화를 살펴보았는데, 이 경우 이중추출율은 10%로 고정시켰으며 X의 범주에 대한 주변합의 비율을 0.5로 하였다. 또 X의 각 범주에 대한 Y의 각 범주의 조건부확률을 0.5로 하였으며 X의 첫번째 범주에만 오분류율 D_1을 1%부터 3%까지 가정하였다($D_2 = 0$으로 고정).

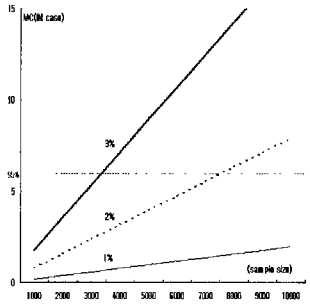

그림 6.1: IN 경우

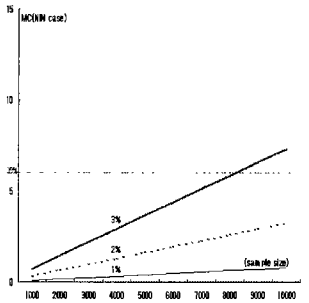

그림 6.2: NIN 경우

그림 6.1은 불확실 분할표의 주변합이 무시할 수 있는 무응답(IN)인 경우에 MC-통계량의 변화를 살펴본 것이다. MC-통계량은 2×2분할표에서 자유도 2인 χ^2-분포를 따르므로 유의수준 10%에서 오분류가 없다는 귀무가설을 기각하기 위해서는 $\chi^2_{(2,0.1)} = 4.61$ 이상의 값을 가져야 한다. 마찬가지로 유의수준 5%에 대해서는 MC-통계량이 $\chi^2_{(2,0.05)} = 5.99$ 이상일 때 귀무가설을 기각할 수 있다. 따라서 원분할표와 완비분할표의 조건부확률의 오분류율이 X의 한 범주에 대해서만 D_1=2%인 경우에는 초기표본이 5,000개 이상이면 유의수준 10%에서 유의하게 나타나며, 7,000개 이상의 초기표본에 대해서는 유의수준 5%에서도 오분류가 없다는 귀무가설을 기각시킬 수 있다. 오분류율 D_1이 3%인 경우는 3,000개 정도의 초기표본으로 유의수준 5%에서 오분류 판정이 가능함을 인지한다. 반면에 오분류율 D_1이 1%인 경우에는 초기표본 10,000에서도 유의수준 10%로 귀무가설을 기각시키지 못함을 인지한다.

그림 6.2는 무시할 수 없는 무응답(NIN)의 경우로 표본추출비용에 대한 가중값을 3/4로 하고 MC-통계량의 변화에 대한 결과이다. 유의수준 5%에서 오분류율 D_1이 2%이하인 경

우에는 귀무가설을 기각할 수가 없다. 만약 오분류율이 D_1=3%이라면 초기표본이 9,000개인 경우에 MC-통계량의 값은 6.575이며 p-값은 0.037으로 유의수준 5%로 오분류 판정을 할 수가 있다.

6.2. 이중추출율 변화에 따른 MC-통계량

이중추출율을 1%에서 30%까지 변화시키면서 MC-통계량의 변화를 살펴보았는데, 이 경우에는 원분할표의 표본크기를 10,000개로 고정시키고, X의 범주에 대한 주변합의 비율을 0.5로 하였다. 또 X의 각 범주에 대한 Y의 각 범주의 조건부확률을 0.5로 하였으며 X의 첫번째 범주에만 오분류율 D_1을 1%에서 3%까지 가정하였다($D_2 = 0$으로 고정).

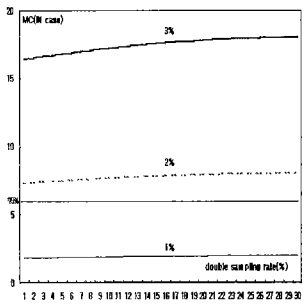

그림 6.3: IN 경우

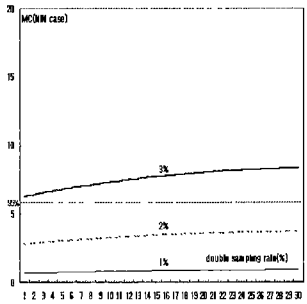

그림 6.4: NIN 경우

그림 6.3은 무시할 수 있는 무응답(IN)의 경우로 오분류율 D_1이 1%인 경우에는 이중추출율이 30%일 때에도 유의수준 5%에서 오분류가 없다는 귀무가설을 기각할 수가 없다. 반면에 D_1이 2%와 3%인 경우에는 이중추출율이 1%에서도 유의수준 5%로 귀무가설을 기각할 수 있다. 또한 이중추출율의 변화에 MC-통계량이 크게 변화하지 않으므로 오분류 검정을 위해서는 이중추출율의 증가보다는 초기표본수가 중요한 요소임을 인지 할 수 있다.

그림 6.4는 무시할 수 없는 무응답(NIN)인 경우이다. 초기표본과 이중추출표본의 단위당 표본조사 비용에 대한 가중값을 3/4으로 하였으며 사전정보 α^*가 없다고 가정하여 경험적 사전모수를 이용하였다. 이 경우에도 무시할 수 없는 무응답의 경우와 동일하게 오분류율이 D_1=3%인 경우에만 유의수준 5%에서 오분류가 없다는 귀무가설을 기각할 수 있으며 오분류율이 1%, 2%인 경우에는 귀무가설을 기각할 수가 없다.

6.3. 주변합의 변화에 따른 MC-통계량

표본크기는 10,000개, 이중추출율은 10%로 고정시키고 X의 각 범주에 주변합에 대한 Y의 첫번째 범주에 조건부확률을 0.5로 설정하였다. 오분류율이 D_1=2%와 D_2=0%일 경우에 원분할표에서 X의 첫번째 범주에 대한 주변합의 비율을 0.1에서 0.9로 증가시키면서 MC-통계량의 변화를 살펴보았다.

표 6.1에서 무시할 수 있는 무응답(IN)인 경우는 주변합의 비율을 0.1에서 0.9로 증가시킬수록 오분류율 D_1=2%에 대해서 MC-통계량의 p-값은 0.453에서 0.0008로 감소하였다.

표 6.1: 주변합의 변화에 따른 MC-통계량

X의 첫번째 범주에 주변합비율	D_1	MC			
		IN		NIN	
		통계량	p-값	통계량	p-값
0.1	2%	1.5858	0.45253	0.2129	0.89903
0.3		4.7561	0.09273	1.5231	0.46694
0.5		7.9264	0.01900	3.2459	0.19732
0.7		11.0966	0.00389	5.1106	0.07767
0.9		14.2669	0.00080	7.0401	0.02960

무시할 수 없는 무응답(NIN)의 경우도 주변합의 비율이 증가할수록 MC-통계량의 p-값은 0.899에서 0.03으로 감소하고 있으므로, 주변합의 비율이 증가할수록 오분류율에 대한 검정력이 증가하는 것을 파악할 수 있다.

또한 무시할 수 있는 무응답과 무시할 수 없는 무응답의 경우에 MC-통계량의 p-값들은 많은 차이를 보이고 있다. 이것은 무시할 수 있는 무응답의 경우는 완비분할표의 정보만을 고려하였기 때문에 베이지안 추정과정에서 오분류율 D_1=2%가 모두 반영된 결과이며 무시할 수 없는 무응답의 경우는 완비분할표와 불확실 분할표의 정보를 표본조사비용에 의해 가중하여 고려하였기 때문에 오분류율이 어느 정도 상쇄되어 반영되었기 때문이다. 따라서 무시할 수 있는 무응답인 경우보다 무시할 수 없는 무응답인 경우가 오분류가 없다는 귀무가설을 기각하기에 더욱 보수적(conservative)임을 인식할 수 있다.

6.4. 조건부확률의 변화에 따른 MC-통계량

표본크기는 10,000개로 고정시키고 이중추출율은 10%, X의 범주에 주변합비율을 0.5, 오분류율은 D_1=1%와 D_2=0%로 고정시켰다. 원분할표의 X의 각 범주에 서 Y의 첫번째 범주에 대한 조건부확률을 0.1에서 0.9로 증가시키면서 MC-통계량의 변화를 살펴보았다.

표 6.2: 조건부확률의 변화에 따른 MC-통계량

Y의 첫번째 범주에 조건부확률	D_1	MC			
		IN		NIN	
		통계량	p-값	통계량	p-값
0.1, 0.9	1%	5.2915	0.071	2.1986	0.333
0.3, 0.7		2.3378	0.311	0.9607	0.619
0.5		1.9805	0.371	0.8113	0.666

표 6.2에서 무시할 수 있는 무응답(IN)과 무시할 수 없는 무응답(NIN)의 경우 모두 조건부확률이 0이나 1에 가까울수록 MC-통계량에 대한 p-값이 0.371에서 0.071로, 0.666에서 0.333으로 감소하여 검정력이 증가하는 것으로 파악할 수 있다. 또한 조건부확률이 0.1과 0.9인 경우와 0.3과 0.7인 경우는 동일한 MC-통계량의 값을 갖는데 이는 2×2분할표에서는 X의 각 범주에 대해서 Y의 범주에 대한 칸값은 이항형태로 나타나기 때문이며 조건부확률이 0이나 1에 가까워지면 MC-통계량의 값이 급격히 증가함을 알 수 있다. 이는 조건부 확률의 극값, 즉 칸 도수가 적을 수록 MC-통계량이 민감하게 반응한다는 것을 인지한다.

6.5. 범주수준별 검정통계량

2×2분할표에서 MC-통계량이나 Pearson X^2-통계량은 자유도 2인 χ^2-분포를 따르며 분할표 전체의 오분류 여부를 검정하는 통계량이다. 만약 오분류가 없다고 고려되는 변수의 각 범주별 오분류 검정을 위해서는 (5.3)식에서 제시하고 있는 통계량 $MC_i(\theta_{j|i})$를 이용하여 검정할 수 있는데 이때의 검정통계량은 자유도가 1인 χ^2-분포를 따른다. 표 6.3은 표본크기가 10,000개이고 이중추출율이 10%이며 X의 첫번째 범주의 주변합비율이 0.5, X의 각 범주에서 Y의 첫번째 범주에 대한 조건부확률을 0.5로 고정시키고 오분류율 D_1과 D_2를 동시에 고려하여 X의 수준별로 오분류 검정을 실시하였다.

표 6.3: 범주수준별 검정통계량

	오분류율		MC_1		MC_2		MC	
	D_1	D_2	통계량	p-값	통계량	p-값	통계량	p-값
IN	1%	2%	1.981	0.159	7.926	0.004	9.907	0.007
NIN			0.811	0.368	3.246	0.072	4.057	0.312

표 6.3에서 무시할 수 있는 무응답(IN)의 경우에 MC-통계량에 대한 p-값은 0.007로 유의수준 1%에서도 유의한 결과를 나타내고 있다. 그러나 범주별 통계량을 살펴보면 MC_1-통계량에 대한 p-값은 0.159로 유의수준 10%에서도 유의하지 않으나 MC_2-통계량의 p-값은 0.004로 유의수준 1%에서도 유의한 결과를 나타내고 있다. 따라서 분할표 전체에 오분류가 존재한다고 판단할 수 있으며 더욱 정확하게 X의 첫번째 범주에서는 잘 분류되어있는 반면에 두번째 범주에 오분류가 존재한다고 결론을 내릴 수 있다.

무시할 수 없는 무응답(NIN)의 경우를 살펴보면 MC-통계량에 대한 p-값은 0.132로 귀무가설을 기각할 수가 없으므로 전반적으로 잘 분류된 분할표자료라고 판단하기 쉽다. 그러나 범주별 통계량을 살펴보면 MC_1-통계량에 대한 p-값은 0.368로 유의하지 않으나 MC_2-통계량의 p-값은 0.072로 유의수준 10%에서 귀무가설을 기각한다. 이는 분할표에서 X의 두번째 범주가 오분류되어 있다는 사실을 유도하였다. MC-통계량과 MC_i-통계량을 이용하면 분할표 전체의 오분류뿐만 아니라 각 범주별로도 오분류에 대한 검정이 가능한 장점이 있다.

7. 결론

본 논문에서는 범주형 자료의 오분류 여부를 검정하는 방법으로 무응답값을 추정하는 방법 을 이용하였는데 $I \times J$ 분할표에서 오분류 가능성이 없는 변수에 대한 주변합은 고정시키고, 오분류 가능성이 있는 변수의 주변합을 무응답으로 고려하고 무응답값의 추정하여 초기자료의 오분류 여부를 검정하였다. 무응답의 추정은 경험적인 베이지안 추정법을 이용하였는데 Sebastiani와 Ramoni(1997)가 제안한 Bound와 Collapse의 개념을 이용하였고 사전모수에 대한 분포로 Dirichlet분포를 고려하여 다음과 같은 기대 베이지안 추정량을 구하였다(식(3.7) 참조).

$$\hat{\theta}_{j|i}|S_{est} = \frac{\alpha_{ij} + n_{ij} + \omega \cdot \phi_{j|i} N^*_{i+}}{\alpha_{i+} + n_{i+} + \omega \cdot N^*_{i+}}.$$

추정을 위한 정보로는 Tenenbein(1970)에 의해 제안된 이중추출자료와 표본조사비용을 고려하였는데 무시할 수 없는 무응답의 경우에 외부정보는 다음과 같이 정의하였다(식(4.2) 참조).

$$\phi_{j|i} = \frac{c_1}{C}\frac{\alpha_{ij} + n_{ij}}{\alpha_{i+} + n_{i+}} + \frac{c_2}{C}\frac{\alpha_{ij} + N^*_{ij}}{\alpha_{i+} + N^*_{i+}}.$$

사전분포에 대한 사전모수는 타 연구기관의 조사결과나 혹은 몇 년 전에 실시된 유사한 조사와 같은 사전정보 등을 고려하였으며 사전정보가 없을 경우에는 완비표본과 불확실 표본의 정보를 고려하여 사전모수를 구하였다(식(4.3) 참조).

$$\alpha_{ij} = n\left[k\left(\frac{c_1}{C}\frac{n_{ij}}{n} + \frac{c_2}{C}\frac{N^*_{ij}}{N^*}\right) + (1-k)\alpha^*_{ij}\right].$$

추정된 경험적 베이지안 추정값을 이용하여 오분류의 정도파악하기 위해 제안된 MC-통계량은 오분류가 없는 변수에 대한 범주별 오분류 검정통계량의 합으로 자유도가 $I \times (J-1)$인 χ^2분포를 따르며 다음과 같이 표현되는데

$$MC(\theta_{j|i}) = \sum_{i=1}^{I} MC_i(\theta_{j|i}), \quad i = 1, \ldots, I, \; j = 1, \ldots, J$$

각 범주수준에 따라 오분류 검정할 수 있다는 특징이 있다.

참고문헌

[1] Barron, B.A. (1977). "The effects of misclassification on the estimation of relative risk," *Biometrics*, Vol. 33, pp. 414-418.

[2] Bross, I. (1954). "Misclassification in tables," *Biometrics*, Vol. 10, pp. 478-486.

[3] Fuchs, C. (1982). "Maximum likelihood estimation and model selection in contingency tables with missing data," *Journal of the American Statistical Association*, Vol. 77, pp. 270-278.

[4] Good, I.J. (1968). "*The estimation of probability: An essay on modern bayesian methods*," MIT press, Cambridge.

[5] Hochberg, Y. (1977). "On the use of double sampling schemes in analyzing categorical data with misclassification errors," *Journal of the American Statistical Association*, Vol. 72, pp. 914-921.

[6] Kalton, G. (1983). "*Compensating for missing survey data*," Research report series, Institute for social research.

[7] Little, R.J.A. and Rubin, D.B. (1987). "*Statistical analysis with missing data*," John Wiley & Sons, New York.

[8] Park, T.S. and Brown, M.B. (1994). "Models for categorical data with nonignorable nonresponse," *Journal of the American Statistical Association*, Vol. 89, No. 45, pp. 44-52.

[9] Sebastiani, P. and Ramoni, M. (1997). "Bayesian inference with missing data using bound and collapse," *KMI-TR., No.* 58, Open University.

[10] Tebaldi, C. and West, M. (1998). "Reconstruction of contingency tables with missing data," *Duke technical report.* Duke University.

[11] Tenenbein, A. (1970). "A double sampling scheme for estimation from binomial data with misclassification," *Journal of the American Statistical Association*, Vol. 65, pp. 1350-1361.

[12] Tenenbein. A. (1971). "A double sampling scheme for estimation from binomial data with misclassification: sample size determination," *Biometrics*, Vol. 27, pp. 935-944.

[2000년 5월 접수, 2000년 11월 채택]

Empirical Bayesian Misclassification Analysis on Categorical Data

Han Seung Lim[1] Chong Sun Hong[2] Munsup Seoh[3]

ABSTRACT

Categorical data has sometimes misclassification errors. If this data will be analyzed, then estimated cell probabilities could be biased and the standard Pearson X^2 tests may have inflated true type I error rates. On the other hand, if we regard well-classified data with misclassified one, then we might spend lots of cost and time on adjustment of misclassification. It is a necessary and important step to ask whether categorical data is misclassified before analyzing data. In this paper, when data is misclassified at one of two variables for two-dimensional contingency table and marginal sums of a well-classified variable are fixed. We explore to partition marginal sums into each cells via the concepts of Bound and Collapse of Sebastiani and Ramoni (1997). The double sampling scheme (Tenenbein 1970) is used to obtain informations of misclassification. We propose test statistics in order to solve misclassification problems and examine behaviors of the statistics by simulation studies.

Keywords: Double sampling scheme; Empirical Bayesian estimation; External-informati -on; MC-statistic; Misclassification; Missing data; Posterior estimator; Pre-information; Prior parameter.

1) Manager, RMS Department, Credit Rating Division, Korea Management Consulting & Credit Rating Corporation. E-mail: haslim@kmcc.com
2) Professor, Department of Statistics, SungKyunKwan University. E-mail: cshong@skku.ac.kr
3) Professor, Department of Mathematics and Statistics, Wright State University. E-mail: munsup.seoh@wright.edu