한국통계학회논문집 제4권 1호
1997년 4월 pp. 75-90

반복조사를 통한 범주형 자료의 오분류 탐색[1)]

고 봉 성[2)]

요 약

본 연구는 범주형자료의 오분류에 관한 연구로, 2×2분할표의 자료에 오분류가 있다고 생각되는 조사와 반복조사를 통해 정확하게 분류한 새로운 범주형자료를 시간이라는 새 변수의 결합을 통해 오분류 여부를 탐색하는 방법에 대한 연구이다.

1. 서 론

두 변수가 교차 분류된 범주형자료의 분석은 피어슨(Pearson) X^2검정과 우도비 G^2을 통하여 변수들간의 독립성검정이나 동질성검정에 대한 분석을 하여왔다. 이와 같은 연구로 Chiacchierini와 Arnold(1977), Hochberg(1977) 등은 2×2분할표로 구성된 Tenenbein(1970)의 연구를 다차원으로 확장하고 이에 대한 독립성검정 등을 수행하였다. 그러나 삼차원이상으로 교차 분류된 다차원 분할표는 여러 가지 복잡한 구조를 가질 수 있기 때문에 구조적 특성의 탐색을 위해서는 카이제곱 검정이상의 체계적인 통계적 방법이 필요하며, 이러한 통계적 분석을 위한 모형이 대수선형모형(log linear model)이다. 이와 같은 대수선형모형은 모형에 포함되는 효과들을 살펴봄으로써 변수들간의 연관성을 파악하는 측도가 되기 때문에 범주형 자료가 가지고 있는 변수들간의 관계를 파악하는 데 유용한 도구가 되며, Chen(1979)은 이중추출법을 이용하여 오분류가 있는 범주형자료의 검정과 모형선택을 위해 대수선형모형을 이용하였고, Hochberg와 Tenenbein(1983), Chen과 Hochberg 그리고 Tenenbein(1984)등은 이중추출법을 확장한 삼중추출법(triple sampling)을 이용한 오분류의 추정을, Espeland와 Odoroff(1985)는 대수선형모형의 최우추정값을 얻기 위해 EM알고리즘을 이용한 χ^2검정 등을 수행하였다.

한편, 본 연구에서는 2×2분할표로 분류된 사전조사의 오분류 여부를 파악하기 위한 방법으로, 시간이라는 매개변수와의 결합을 통한 반복조사에 의해 새로운 2×2×2분할표로 구성된 이용하여 사전조사의 오분류를 탐색하는 방법에 대해 연구하였다.

1) 이 논문은 1996년도 전주대학교 학술연구비 지원에 의해 연구되었습니다.
2) (560-759) 전북 전주시 완산구 효자동 3가 1200, 전주대학교 응용통계학과 전임강사

2. 반복조사의 오분류

어떤 시점에서 특정한 수의 표본으로 부터 수집된 자료는 각각 두 개의 범주를 가지는 변수 1과 변수 2로 분류된다고 하면, 이는 다음과 같은 2×2분할표로 나타낼 수 있다.

구분		변수 2		합
		$j=1$	$j=2$	
변수 1	$i=1$	x_{11}	x_{12}	x_{1+}
	$i=2$	x_{21}	x_{22}	x_{2+}
합		x_{+1}	x_{+2}	N

이 표에서 x_{ij}는 각 칸의 관측도수이며, 이를 각 칸의 확률에 자연대수를 취한 대수선형모형으로 나타내면 다음과 같다.

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

여기서 $u_{12(ij)}$는 변수 1과 변수 2의 연관항을 나타내며, $u_{1(i)}$, $u_{2(j)}$항은 주효과항을, u는 전체 평균을 나타낸다.

한편, 일정 시간의 경과후 사전조사의 오분류에 관심을 갖게 된다면 이는 앞과 동일한 변수의 범주로 구성된 2×2분할표에 대한 자료를 얻게 된다.

즉, N_1개의 초기조사는 오분류 가능성이 있는 자료이며 새로운 두 번째 조사의 자료는 초기의 조사와는 표본 크기가 다른 오분류 없이 정확하게 분류된 N_2개의 자료이다.

이와같은 자료에서 N_1개의 이차원분할표와 새로운 N_2개의 2×2 분할표를 시간이라는 새로운 변수의 범주 k로 간주할 수 있으며, 이 자료구조는 다음과 같은 2×2×2인 삼차원분할표로 정리할 수 있다.

<표 1> 결합된 2×2×2분할표

구분				변수 2		합
				$j=1$	$j=2$	
변수 3	$k=1$	변수 1	$i=1$	x_{111}	x_{121}	x_{1+1}
			$i=2$	x_{211}	x_{221}	x_{2+1}
		합		x_{+11}	x_{+21}	$x_{++1}=N_1$
	$k=2$	변수 1	$i=1$	x_{112}	x_{122}	x_{1+2}
			$i=2$	x_{212}	x_{222}	x_{2+2}
		합		x_{+12}	x_{+22}	$x_{++2}=N_2$

이때 두 개의 이차원분할표에 대한 새로운 세 번째 변수 k의 각 범주수준은 다음과 같은 새로운 대수선형모형으로 정의되며

$$\log p_{ijk} = v^{(k)} + v^{(k)}_{1(i)} + v^{(k)}_{(2)} + v^{(k)}_{12(ij)}, \qquad k=1,2\ ,$$

새 변수 k에 대한 주효과와 교호작용항은 전체 평균

$$u=\frac{1}{k}\sum_{k} v^{(k)}$$

와 변수 1과 2에 대한 주효과와 교호작용항으로 나타난다.

$$u_{1(i)}=\frac{1}{k}\sum_{k} v_{1(i)}^{(k)}\ ,\quad u_{2(j)}=\frac{1}{k}\sum_{k} v_{2(j)}^{(k)}\ ,\quad u_{12(ij)}=\frac{1}{k}\sum_{k} v_{12(ij)}^{(k)}\ ,$$

여기서 $u_{12(ij)}$는 변수 1과 변수 2의 상호작용의 평균으로 u_{12}는 부분 연관항 즉, 세 번째 변수의 효과를 무시한 변수 1과 2의 부분적 연관을 의미한다. 그리고 이들 세 번째 변수의 평균을 나타내는 효과들에 의한 편차에 의해 다음과 같이 변수 3에 대한 주효과와 교호작용을 정의할 수 있다.

$$u_{3(k)}=v^{(k)}-u\ ,\quad u_{13(ik)}=v_{1(i)}^{(k)}-u_{1(i)}\ ,\quad u_{23(jk)}=v_{2(j)}^{(k)}-u_{2(j)}\ ,$$

$$u_{123(ijk)}=v_{12(ij)}^{(k)}-u_{12(ij)}\ .$$

한편, <표 1>은 각기 다른 두 집단에서의 분할표를 의미하게 되며, 각 표본추출상태에 따라 합이 N_1과 N_2로 고정된 적다항모형(product-multinomial)을 따르게 된다.

이때 각 표본추출상태에서의 확률을 살펴보면, $k=1$인 경우의 확률 p_{ij1}는

$$p_{ij1}=\frac{x_{ij1}}{N_1}$$

으로 나타나며, $k=2$인 경우는

$$p_{ij2}=\frac{x_{ij2}}{N_2}$$

로 나타난다.

이때 각 k층에서의 상대비율에 관심을 두게 되면 변수 1, 2의 수준이 주어진 경우 변수 3의 첫 번째 수준에 속할 조건부확률은 다음과 같이 구할 수 있다.

$$\begin{aligned} p_{1(ij)} &= \Pr(\text{변수 3의 첫번째수준} \mid \text{변수 1의 } i\text{번째수준, 변수 2의 } j\text{번째수준}) \\ &= p_{ij1}/(p_{ij1}+p_{ij2}) \end{aligned}$$

여기서 $i=1,2$이며 $j=1,2$이다. 그리고 변수 1과 2의 수준이 주어진 경우에 변수 3의 두 번째 수준에 대한 확률은 $p_{2(ij)}=p_{ij2}/(p_{ij1}+p_{ij2})$와 같으며, 변수 3의 수준에 대한 상대비율은

$$\frac{p_{1(ij)}}{p_{2(ij)}}=\frac{p_{1(ij)}}{1-p_{1(ij)}}=\frac{m_{ij1}}{m_{ij2}}$$

와 같게 되며, 이에 자연대수를 취하면 변수 1과 2의 수준이 주어졌을 때 표본추출상태의 차로 나타나는데, 이는 일반적인 이항변수의 로지트변환과 같게 된다. 따라서 <표 1>의 모형에서 상대비율에 대한 로지트 변환은 변수 1과 2가 변수 3에 미치는 영향을 연구하기 위한 방법으로 적용될 수 있다.

한편, 삼차원 포화모형

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \tag{1}$$

에서 변수 3의 k 수준에 대한 상대비율은 다음과 같게 된다.

$$\begin{aligned} \log\left(\frac{p_{ij1}}{p_{ij2}}\right) &= \log p_{ij1} - \log p_{ij2} \\ &= (u_{3(1)} - u_{3(2)}) + (u_{13(i1)} - u_{13(i2)}) \\ &\quad + (u_{23(j1)} - u_{23(j2)}) + (u_{123(ij1)} - u_{123(ij2)}) \\ &= 2\{u_{3(1)} + u_{13(i1)} + u_{23(j1)} + u_{123(ij1)}\} \end{aligned} \tag{2}$$

$$\equiv w + w_{1(i)} + w_{2(j)} + w_{12(ij)} \tag{3}$$

즉, 표본추출상태에 따른 상대비율은 변수 3의 주효과, 변수 1과 3, 변수 2와 3의 교호작용항에 의존함을 알 수 있으며, 전체적으로는 변수 3을 포함한 교호작용항의 두배 합으로 나타나게 된다.

한편, 삼차원포화모형에서의 제약조건을 다음과 같이 Z을

$$z(l) = \begin{cases} 1 & , \text{ 만약 } l=1 \\ -1 & , \text{ 만약 } l=2 \end{cases}$$

$$z(m, n) = \begin{cases} 1 & , \text{ 만약 } m=1,\ n=1 \\ -1 & , \text{ 만약 } m=1,\ n=2 \\ -1 & , \text{ 만약 } m=2,\ n=1 \\ 1 & , \text{ 만약 } m=2,\ n=2 \end{cases}$$

으로 정의하면, 식 (3)은 다음과 같이 표현될 수 있다.

$$\text{logit(ij)} = z + z_{1(i)} + z_{2(j)} + z_{12(i,j)} \tag{4}$$

따라서 위 모형은 변수 1과 2의 수준조합이 표본추출상태에 미치는 영향을 알아내기 위한 모형이 됨을 알 수 있다.

가정을 간단하기 위해 오분류 가능성이 있는 표본 N_1, N_2에서 변수 2는 남자와 여자같이 정확하게 분류되는 범주라 한다면, 표본 N_1에서의 오분류는 결국 변수 1의 i수준에 따라 변하게 된다. 따라서 2×2×2분할표에서 오분류의 판단은 변수 1과 2의 수준조합이 변수 3에 의해 좌우되는 것을 알 수 있으며, 이를 판단하기 위해 (1)과 같은 삼차원 분할표의 대수선형모형을 고려할 수 있다. 먼저 표본추출상태를 의미하는 변수 3이 포함되는 모형들을 고려해 보면 다음과 같다.

[13][2], [23][1], [12][13], [12][23], [13][23], [123]

여기서 모형 [123]은 변수 1,2,3의 교호작용항이 모두 포함된 삼차원분할표의 포화모형을 의미하며, 모형 [13][2]은 삼차원 포화모형에서 연관항 u_{123}, u_{12}, u_{23}이 제외된 조건부 독립 모형을 의미한다.

만약, 주어진 자료가 삼차원 포화모형을 따른다고 가정하면, 이때 변수 3의 수준 조합에 대한 차는 식 (3)과 같이 나타나게 되며, 관심을 갖게 되는 $w_{12(ij)}$항의 추정값에 의해 오분류를 파악할수 있는 방법으로 제시될 수 있다. 여기서 $w_{12(ij)}$항은 오분류가 있다고 생각되는 표본과 정확하게 분류된 표본과의 승산비(odds-ratio)의 차로 나타나게 되는데, 이때 오분류 자료의 영향을 살펴

보면 다음과 같다. 먼저 오분류 가능성이 없이 정확하게 분류된 각 칸의 확률을 p_{ij2}라 하고 i와 j는 두 개의 범주로 구성된다고 하자. 그리고 오분류 가능성이 있는 표본에서의 칸 확률들은 p'_{ij2} 만큼의 오분류가 있다고 생각하면 오분류 가능성이 있는 각 칸의 확률은 정확하게 분류된 범주의 칸 확률에다 오분류 정도를 합한 형태로 생각할 수 있다. 이와 같은 가정하에 표본추출상태를 세 번째 변수로 간주하여 표를 작성하면 다음과 같게 된다.

<표 2> 2×2×2분할표의 오분류 가능성

구 분		정확한 분류 변수		합
		$j=1$	$j=2$	
오분류 가능성이 있는 표본 $k=1$	$i=1$	$p_{112}+p'_{112}$	$p_{122}+p'_{122}$	$p_{1+2}+p'_{1+2}$
	$i=2$	$p_{212}+p'_{212}$	$p_{222}+p'_{222}$	$p_{2+2}+p'_{2+2}$
	합	$p_{+12}+p'_{+12}$	$p_{+22}+p'_{+22}$	
정확한 분류 방법에 의한 표본 $k=2$	$i=1$	p_{112}	p_{122}	p_{1+2}
	$i=2$	p_{212}	p_{222}	p_{2+2}
	합	p_{+12}	p_{+22}	

<표 2>에서 변수 2는 정확한 분류 변수에 의해 구성된 범주이므로 변수 2의 j번째 범주의 칸 확률의 합은 일정하게 되며, 확률은 음수가 아니므로 다음과 같은 조건들을 생각할 수 있다.

$$p'_{112}=p'_{212}=0\ ,\ p'_{122}=p'_{222}=0\ ,\ p_{+12}+p_{+22}=1$$

이때 주어진 자료에 대한 오분류 차이는 앞에서 논의한 바와 같이 $w_{12(ij)}$항에 의해 묘사되게 되며 이는 다음과 같이 나타난다.

$$\begin{aligned} w_{12(11)} &= u_{123(111)}-u_{123(112)} \\ &= 2u_{123(111)} \end{aligned}$$

여기서 삼차원 포화모형의 연관항 $u_{123(ijk)}$는 다음과 같이 나타난다.

$$u_{123(ijk)} = \frac{1}{8}\log\left\{\frac{\alpha^{(1)}}{\alpha^{(2)}}\right\},$$

여기서 $\alpha^{(k)}=\dfrac{p_{11k}p_{22k}}{p_{12k}p_{21k}}$ 이다. 따라서

$$w_{12(ij)}=\frac{1}{8}\log\left\{\frac{(p_{112}+p'_{112})(p_{222}+p'_{222})}{(p_{122}+p'_{122})(p_{212}+p'_{212})}\right\}-\frac{1}{8}\log\left\{\frac{p_{112}p_{222}}{p_{122}p_{222}}\right\} \tag{5}$$

로 나타나게 된다.

만약 오분류가능성이 있는 사전조사의 범주의 확률 $p_{112}+p'_{112}$가 정확한 분류방법에 의한 표본

의 범주의 확률 p_{+12}와 같고 사전조사의 범주의 확률 $p_{122}+p_{122}$와 p_{122}와 같다면 변수 1의 행들 간에는 오분류가 없다는 의미가 된다. 이와 같은 결과는 역으로 식 (5)에서 오분류가 가능성이 있는 사전조사와 정확한 분류방법에 의한 반복조사와의 오분류를 나타내준다.

3. 오분류 모형의 추정과 검정

모형 (2)의 최우추정을 위해서는 우도방정식을 구해야 한다. 그런데 로지트는 오분류 가능성이 있는 자료에 의해서 관측된 x_{ij1}와 정확한 자료에 의해 분류된 x_{ij2}가 주어진 $x_{ij+}=x_{ij1}+x_{ij2}$이며, 이는 x_{ij1}의 분포의 모수이므로 x_{ij1}는 다음과 같은 이항분포를 따른다.

$$\Pr(X_{ij1}=x_{ij1}) = \binom{x_{ij+}}{x_{ij1}} p_{1|ij}^{\,x_{ij1}}(1-p_{1|ij})^{x_{ij2}} \quad ,$$

여기서

$$p_{1|ij}=\frac{p_{ij1}}{p_{ij1}+p_{ij2}} = \frac{\exp\{\ \mathrm{logit}\,(ij)\}}{\{1+\exp[\ \mathrm{logit}\,(ij)]\}}$$

이다.

따라서 $\{X_{ij1}\}$의 결합분포는 이항분포의 곱으로 다음과 같은 적이항분포(product binomial distribution)를 따르게 된다. 그러므로 우도함수는

$$L(w, w_1, w_2, w_{12}) = \prod_{i=1}^{2}\prod_{j=1}^{2}\binom{x_{ij+}}{x_{ij1}} p_{1|ij}^{\,x_{ij1}}(1-p_{1|ij})^{x_{ij2}}$$

이 된다. 그리고 대수우도함수는 식 (4)로부터 다음과 같게 된다.

$$\begin{aligned}\log L(\boldsymbol{w}) &= \sum_{i=1}^{2}\sum_{j=1}^{2} x_{ij1}\{w+w_1 z(i)+w_2 z(j)+w_{12} z(i,j)\} \\ &\quad -\sum_{i=1}^{2}\sum_{j=1}^{2}\log\{1+\exp[\ \mathrm{logit}\,(ij)]\} \ + \ \log\left\{\binom{x_{ij+}}{x_{ij1}}\right\} \\ &= x_{++1}w+w_1(x_{1+1}-x_{2+1}) + w_2(x_{+11}-x_{+21}) \\ &\quad + w_{12}(x_{111}-x_{121}-x_{211}+x_{221}) \\ &\quad -\sum_{i=1}^{2}\sum_{j=1}^{2}\log\{1+\exp[\ \mathrm{logit}\,(ij)]\} \ + \ \log\left\{\binom{x_{ij+}}{x_{ij1}}\right\}\end{aligned}$$

따라서 $\boldsymbol{w}$항의 최우추정을 위한 우도방정식은 다음과 같이 얻을 수 있다.

$$\hat{m}^{*}_{++1} = x_{++1} \tag{6}$$

$$\hat{m}^{*}_{1+1}-\hat{m}^{*}_{2+1} = x_{1+1}-x_{2+1} \tag{7}$$

$$\hat{m}^{*}_{+11}-\hat{m}^{*}_{+21} = x_{+11}-x_{+21} \tag{8}$$

$$\hat{m}^{*}_{111}-\hat{m}^{*}_{121}-\hat{m}^{*}_{211}+\hat{m}^{*}_{221} = x_{111}-x_{121}-x_{211}+x_{221} \quad , \tag{9}$$

여기서 $\hat{m}^{*}_{ij1} = E[x_{ij1}|x_{ij+}] = x_{ij+}\, p_{1|ij}$ 이다. 이때 식(6)에서 부터 식(9)까지의 관계에 의해

$$\hat{m}^{*}_{1+1} = x_{1+1} \tag{10}$$

$$\hat{m}^{*}_{+11} = x_{+11} \tag{11}$$

를 얻을 수 있고 그리고

$$\hat{m}^{*}_{111} = x_{111} \tag{12}$$

을 구할 수 있다. 그런데 만약

$$n(p_{ij1}+p_{ij2}) = np_{ij+} = x_{ij+} \tag{13}$$

라고 한다면, 식 (6)은 다음과 같은 등식을 만족하게 된다.

$$\begin{aligned}\sum_i\sum_i x_{ij+}\frac{p_{ij1}}{p_{ij1}+p_{ij2}} &= \sum_i\sum_j np_{ij1}\\ &= np_{++1}\\ &= x_{++1}\end{aligned} \tag{14}$$

마찬가지로 식(10)부터 식(12)는 다음의 등식과 같음을 알 수 있다.

$$np_{1+1}=x_{1+1}\ ,\ np_{+11}=x_{+11}\ ,\ np_{111}=x_{111}\ . \tag{15}$$

즉, 식(14)와 식(15)는 대수선형모형의 u, $u_{13(11)}$, $u_{23(11)}$ 그리고 $u_{123(111)}$의 추정량을 위한 우도방정식이다. 그러나 식(13)이 성립하기 위해서는 $u_{12(ij)}$의 추정을 위한 우도방정식이 만족되어야 한다. 따라서 모형 (4)의 최우추정을 위한 우도방정식 식(6)부터 식(9)는 모형 (1)의 모수들의 최우추정을 위한 모든 우도방정식이 성립되어야만 만족되게 된다.

이러한 필요충분조건이 만족된다면 식(14)와 식(15)로 부터 $u, u_{13(11)}, u_{23(11)}$ 그리고 $u_{123(111)}$의 최우추정량을 얻게 되므로 w, w_1, w_2, w_{12}의 최우추정량은 (2)의 관계로부터

$$\hat{w} = 2\hat{u}_{3(1)}\ ,\ \hat{w}_1 = 2\hat{u}_{13(11)},\ \hat{w}_2 = 2\hat{u}_{23(11)}\ ,\ \hat{w}_{12} = 2\hat{u}_{123(111)}$$

과 같이 구하게 된다.

한편, 앞에서 제시한 삼차원분할표에서 세 번째 변수는 표본추출상태를 나타내므로 주어진 자료에 대한 적합도 검정은 다음과 같은 네가지 모형들을 고려할 수 있다.

가) $\text{logit}(ij) = w + w_1 z(i) + w_2 z(j)$

나) $\text{logit}(ij) = w + w_1 z(i)$

다) $\text{logit}(ij) = w + w_2 z(j)$

라) $\text{logit}(ij) = w$

이들 모형 중에서 먼저 모형 (가)를 위한 적합도검정의 가설은 오분류 가능성이 있는 자료와 정확하게 분류된 자료에 의해 삼차원 분할표로 구성된 모형 (4)식에 w_{12}항이 포함되는가를 가설로 제시하게 된다. 즉,

$$H_0 : w_{12} = 0$$
$$H_1 : w_{12} \neq 0$$

이며, 이의 귀무가설과 대립가설을 다시 쓰면 다음과 같은 의미를 갖는다.

$$H_0 : \text{logit}(ij) = w + w_1 z(i) + w_2 z(j)$$
$$H_1 : \text{logit}(ij) = w + w_1 z(i) + w_2 z(j) + w_{12} z(ij)$$

즉, 위 가설에서 귀무가설이 기각된다면 주어진 삼차원 분할표로 가장 잘 설명하는 모형은 w_{12}항이 포함된 삼차원모형에서의 포화모형이 된다는 의미이며, 채택된다면 (가) 이외의 모형에 의해 적합도 검정을 실시하게 된다.

한편, 앞에서 논의한 w의 분포는 표본비율 p또는 m의 함수로 나타난다. 그런데 w는 표본비율 p의 함수, $w=f(p)$ 이므로 δ 방법을 이용한 점근분포는 다변량정규분포를 따르며, 이의 분산은 다음과 같다.

$$Var(\hat{w}) \simeq \left(\frac{2}{IJK}\right)^2 \sum_{l,m,n} m_{1jk}^{-1} + \left(\frac{4(I-2)}{IJ^2K^2}\right) \sum_{m,n} m_{1jk}^{-1}$$

$$Var(\hat{w}_1) \simeq \left(\frac{4(I-2)(K-2)}{IJK^2}\right) \sum_{\substack{m \neq j \\ l,n}} m_{ijk}^{-1} + \left(\frac{4(I-2)}{IJ^2K^2}\right) \sum_{\substack{m \neq j \\ l,n}} m_{ijk}^{-1}$$
$$+ \left(\frac{4(K-2)}{I^2J^2K}\right) \sum_{\substack{m \neq j \\ l,n}} m_{ijk}^{-1} + \left(\frac{2}{IJK}\right)^2 \sum_{\substack{m \neq j \\ l,n}} m_{ijk}^{-1}$$

$$Var(\hat{w}_2) \simeq \left(\frac{4(J-2)(K-2)}{IJK^2}\right) \sum_{\substack{l \neq i \\ m,n}} m_{ijk}^{-1} + \left(\frac{4(J-2)}{I^2JK^2}\right) \sum_{\substack{l \neq i \\ m,n}} m_{ijk}^{-1}$$
$$+ \left(\frac{4(K-2)}{I^2J^2K}\right) \sum_{\substack{l \neq i \\ m,n}} m_{ijk}^{-1} + \left(\frac{2}{IJK}\right)^2 \sum_{\substack{l \neq i \\ m,n}} m_{ijk}^{-1}$$

$$Var(\hat{w}_{12}) \simeq \left(\frac{4(I-2)(J-2)(K-2)}{IJK}\right) m_{ijk}^{-1} + \left(\frac{4(I-2)(J-2)}{IJK^2}\right) \sum_{l,m} m_{ijk}^{-1}$$
$$+ \left(\frac{4(I-2)(K-2)}{IJ^2K}\right) \sum_{l,n} m_{ijk}^{-1} + \left(\frac{4(J-2)(K-2)}{I^2JK}\right) \sum_{m,n} m_{ijk}^{-1}$$
$$+ \left(\frac{4(I-2)}{IJ^2K^2}\right) \sum_{l} m_{ijk}^{-1} + \left(\frac{4(J-2)}{I^2JK^2}\right) \sum_{m} m_{ijk}^{-1}$$
$$+ \left(\frac{4(K-2)}{I^2J^2K}\right) \sum_{n} m_{ijk}^{-1} + \left(\frac{2}{IJK}\right)^2 \sum_{l,m,n} m_{ijk}^{-1}$$

즉, w항은 다변량정규분포를 따르게 되며, 오분류 가능성이 있는 자료와 오분류 가능성이 없는 자료간의 연관항 w_{12}은 다음과 같이 표준화를 생각 할 수 있다.

$$Z = \frac{\hat{w}_{12}}{\sqrt{Var(\hat{w}_{12})}}$$

4. 시뮬레이션

N_1과 N_2의 표본 크기가 얼마나 될 때 범주의 오분류 여부를 찾아낼 수 있는가를 살펴보기 위해, 각 표본추출상태가 2×2분할표로 구성된 범주형 자료에 대한 유의성 검정을 실시하였다.

이를 위해서

i) N_1과 N_2의 표본 크기 선정

ii) N_1개의 칸 확률과 관측도수의 구성

iii) N_2의 칸 확률과 관측도수의 구성

를 고려하게 된다. 먼저 N_1과 N_2의 크기가 결정되었다면, 다음으로 정확하게 분류된다고 생각되는 N_2의 자료구조에 대한 생각을 할 수 있는데 <표 3>에서 살펴보듯이 우리가 알 수 있는 정보는 확률의 합이 1이라는 정보외에는 알 수 있는 것이 없다. <표 3>에서 최소한 세개의 확률을 알아야만 칸의 확률을 구성할 수 있으므로 p_{112}, p_{+12} , 승산비 α를 임의로 준다면, 정확하게 분류된 자료 N_2에 대한 칸의 확률을 구성할 수 있다.

<표 3> 정확한 분류 방법에 의한 칸 확률($k=2$)

구 분		변수 2		합
		1	2	
변수 1	1	p_{112}	p_{122}	p_{1+2}
	2	p_{212}	p_{222}	p_{2+2}
합		p_{+12}	p_{+22}	$p_{++2}=1$

다음으로 오분류가 있다고 생각되는 자료의 칸 확률은 <표 2>에서 언급한 것처럼 같이 정확하게 분류된 칸 확률에 어느 정도 오분류가 있게 되나, 변수 2는 정확한 분류 변수에 의해 구성되어 있으므로 오분류 확률은 변수 1의 수준에만 영향을 주게 될 것이다. 따라서 변수 1의 범주 1과 2에 일정 부분의 오분류를 가감하게 되면 오분류 가능성이 있는 표본의 확률을 구성할 수 있게 된다.

따라서 N_2, α, p_{112}, p_{+12}이 주어졌다는 가정 아래, 다음의 알고리즘을 이용하여 N_2개 자료로 구성된 분할표의 관측 도수를 구성하고

```
P+22 = 1 - P+12;
P212 = P+12 - P112;
P222 = ( α*P+22*P212)/(P112+ α*P212);
P122 = P+22 - P222;
X112 = N2*P112;
X122 = N2*P122;
X212 = N2*P212;
X222 = N2*P222;
```

오분류 가능성이 있는 N_1의 관측도수는 정확한 변수 2의 범주간에는 오분류가 없으므로 오분류 가능성이 있는 변수 1의 행간의 오분류 확률 A와 B를 이용하여 구성하였다.

```
X111=N1*(P112+A);
X121=N1*(P122-B);
X211=N1*(P212-A);
X221=N1*(P222+B);
```

위의 알고리즘을 이용하여 구성된 칸 값에 의한 분할표들로부터 삼차원 포화모형 변수들의 연관항 u항들을 추정한 후, 앞에서 제시한 변수들간의 연관항 w항들의 추정값을 구하고, 같이 w항을 표준화하여, w_{12}항의 유의성검정을 실시하였다.

삼차원 포화모형에서 w_{12}항의 유의성검정 결과는 <표 4>부터 <표 9>이다. 즉, 사전조사의 결과에 오분류가 있다고 생각되는 표본 N_1에 대한 반복조사로서 정확하게 분류한 표본 N_2의 크기를 사전조사의 표본 N_1의 10%에서 50%까지 증가시키면서 오분류가 얼마나 될때 w_{12}항이 유의하게 나타나는지를 보여주는 표들이다. 한편, 표의 오분류율에서 좌측의 비율은 정확하게 분류된 변수 2의 첫 번째 열에서의 오분류율을 의미하며, 우측의 오분류율은 정확하게 분류된 변수 2의 두 번째 열에서의 오분류율을 의미한다. 따라서 사전조사의 전체 오분류율은 좌측과 우측의 오분류율을 합한 것이다. 이들 표에서 사전조사의 표본에 대해 반복조사의 크기가 50%인 경우들 만을 살펴보면, <표 4>는 사전조사 표본이 100개인 경우로 오분류율이 약 17%(17개 이상)이상일때 삼차원포화모형의 w_{12}항이 유의하게 나타나는 것을 보여주고 있으며, <표 5>는 사전조사의 표본이 300개인 경우로, 오분류율이 10%(30개) 이상일때 w_{12}항이 유의하게 나타내주고 있다. 그리고 <표 6>에서는 사전조사의 표본이 500개인 경우로, 이때는 전체 오분류율이 약 6%(60개)이상일 때 w_{12}항이 유의함을 나타내 주는 결과들이다.

<표 7>에서 <표 9>는 오분류 가능성이 있는 사전조사의 표본이 1000개로 구성되있는 가정아래 승산비를 1부터 3까지 변화시키면서 살펴본 결과이다. 이 결과 정확하게 분류된 반복조사의 승산비가 클 수록 오분류 확률이 작은 것을 더욱 잘 찾아낼 수 있음을 보여주고 있다.

한편, 표의 결과 중 w_{12}항이 유의하지 않게 나타나는 A와 B의 수준조합은 w_{12}항이 0이 아니다라는 해석을 하게 해준다. 즉, $w_{12}=0$이라면 모형자체가 삼차원 포화모형을 따르지 않으므로, w_{12}항을 통한 오분류 여부를 파악할 수 없게 된다. 따라서 앞절에서 논의한대로 모형의 적합도 검정을 통해 적절한 모형을 찾아 오분류 여부를 파악해야 한다.

$$H_0 : \log\left(\frac{p_{ij1}}{p_{ij2}}\right) = w + w_1 + w_2$$

$$H_1 : \log\left(\frac{p_{ij1}}{p_{ij2}}\right) = w + w_1 + w_2 + w_{12}$$

위와 같은 가설하에 주어진 자료에 w_{12}항이 포함되는지 여부를 보기 위해 <표 6>의 내용 중 w_{12}항이 유의하지 않게 나타나는 A, B, N_1, N_2의 수준조합에 대한 모형의 적합여부를 살펴 본

결과가 <표 10>이다.

<표 10>에서 $C-\alpha$ 와 $F-\alpha$는 정확하게 분류된 반복조사의 표본과 오분류 가능성이 있는 사전조사의 표본에 대한 승산비이며, G^2은 우도비 검정통계량을 의미한다. 또 p-값 옆의 '*' 표시는 귀무가설이 5% 유의수준에서 기각된 것을 의미한다.

<표 10>의 결과에서 살펴 볼 수 있듯이 오분류 표본이 500개이고 정확한 표본이 오분류 표본의 10%에서 50%까지 증가시키면서 살펴본 <표 6>에서 w_{12}항이 차이가 없다고 나타나는 수준조합에 대한 적합도검정 결과, w_{12}항이 제외된 모형이 적합한 것으로 나타나고 있으며, <표 6>에서 w_{12}항이 유의하다고 나타나는 결과에 대한 A, B, N_1, N_2 의 수준 조합을 살펴보면 G^2값이 커지고 p-값이 적게 나타나고 있음을 알 수 있으므로 w_{12}항이 포함된 모형이 적합한 것임을 보여준다.

4. 결 론

범주형 자료의 오분류에 관한 연구로는 Bross(1954)의 연구를 시작으로 Tenenbein(1970)의 이중추출법을 이용한 추정, Chen(1979)은 이중추출법을 이용하여 오분류가 있는 범주형자료의 검정과 모형선택을 위해 대수선형모형을 이용한 연구,Hochberg와 Tenenbein(1983), Chen과 Hochberg 그리고 Tenenbein(1984)등은 삼중추출법을 이용한 오분류의 추정을, Espeland와 Odoroff(1985)는 대수선형모형의 최우추정값을 얻기 위해 EM알고리즘을 이용한 χ^2검정 등이 있다.

기존의 연구중 오분류를 파악하는 데 근간이 되어온 이중추출법은 오분류 가능성이 있다고 생각되는 초기표본에서 초기표본보다 적은 수의 부표본을 단순임의추출법으로 뽑은 다음, 추출된 부표본을 정확하게 분류함으로서 오분류에 대한 정보를 얻고자하는 방법이다. 즉, 이 방법은 초기표본의 정보와 부표본의 정보가 교차분류됨으로써, 결국 초기표본의 개체에 대한 정보와 부표본의 개체에 대한 정보를 모두 알아야만 오분류 정도를 파악할 수 있다는 한계를 가지고 있다.

본 연구에서는 2×2분할표로 분류된 사전조사의 오분류 여부를 파악하기 위한 방법으로, 반복조사에 의해 정확하게 분류된 2×2분할표와의 비교를 통해 사전조사의 오분류를 탐색하는 방법에 대해 연구하였다.

즉, 일정시간의 경과후 2×2분할표로 구성된 자료의 오분류를 알고자하는 한 방법으로 사전조사와 동일한 변수의 범주로 구성된 2×2분할표를 얻을 수 있게 된다. 이때 두 번째의 조사는 사전조사에 의한 반복조사로서 오분류 가능성이 없이 정확하게 분류된 자료이고, 반복조사의 표본은 비용상의 문제등으로 사전조사보다 표본크기가 적은 경우를 고려하였다. 이와 같은 자료에서 사전조사의 이차원분할표와 반복조사의 이차원분할표를 시간이라는 변수의 범주 k로 간주할 수 있으며, 이 자료구조는 2×2×2인 삼차원분할표의 대수선형모형을 고려할 수 있게 된다. 즉, 표본추출상태를 나타내는 세 번째 변수인 k에 의해 결합된 2×2×2분할표의 대수선형모형은 오분류 상태와 정확하게 분류된 범주의 상태간의 차로 나타나게 되며, 대수선형모형의 적절한 로지트변환을 통한 모형은 범주의 오분류 여부를 탐색할 수 있는 한 방법으로 제시할 수 있게 된다.

<표 4> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과 ($N_1 = 100$)

표본크기 / 오분류율		$N_1 = 100$				
		N_2				
		10%	20%	30%	40%	50%
0.11	0.01	.4425	.3005	.2249	.1783	.1471
	0.02	.4076	.2643	.1908	.1468	.1181
	0.03	.3742	.2311	.1605	.1196	.0937
	0.04	.3424	.2008	.1339	.0965	.0735
	0.05	.3120	.1732	.1106	.0769	.0569
	0.06	.2830	.1483	.0905	.0606	**.0434**
	0.07	.2554	.1259	.0731	**.0471**	**.0326**
	0.08	.2293	.1058	.0584	**.0361**	**.0241**
	0.09	.2045	.0880	**.0460**	**.0272**	**.0175**
	0.1	.1810	.0724	**.0357**	**.0201**	**.0125**
	0.11	.1590	.0587	**.0272**	**.0146**	**.0087**
	0.12	.1383	**.0469**	**.0204**	**.0104**	**.0059**
	0.13	.1189	**.0368**	**.0149**	**.0072**	**.0039**
	0.14	.1010	**.0283**	**.0107**	**.0049**	**.0025**
	0.15	.0844	**.0213**	**.0074**	**.0032**	**.0016**
	0.16	.0693	**.0156**	**.0050**	**.0020**	**.0010**
	0.17	.0555	**.0110**	**.0032**	**.0012**	**.0006**
	0.18	**.0433**	**.0075**	**.0020**	**.0007**	**.0003**
	0.19	**.0326**	**.0049**	**.0012**	**.0004**	**.0002**

<표 5> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과($N_1 = 300$)

표본크기 / 오분류율		$N_1 = 300$				
		N_2				
		10%	20%	30%	40%	50%
0.06	0.01	.4577	.3152	.2378	.1894	.1566
	0.02	.3969	.2519	.1781	.1344	.1062
	0.03	.3409	.1979	.1302	.0925	.0695
	0.04	.2899	.1525	.0927	.0617	**.0438**
	0.05	.2437	.1152	.0642	**.0397**	**.0265**
	0.06	.2024	.0851	**.0431**	**.0247**	**.0154**
	0.07	.1658	.0613	**.0280**	**.0147**	**.0085**
	0.08	.1338	**.0430**	**.0176**	**.0084**	**.0045**
	0.09	.1062	**.0293**	**.0106**	**.0046**	**.0022**
	0.1	.0827	**.0193**	**.0061**	**.0024**	**.0011**
	0.11	.0630	**.0123**	**.0034**	**.0012**	**.0005**
	0.12	**.0468**	**.0075**	**.0018**	**.0005**	**.0002**
	0.13	**.0337**	**.0043**	**.0009**	**.0002**	**.0001**
	0.14	**.0235**	**.0024**	**.0004**	**.0001**	**.0000**
	0.15	**.0157**	**.0012**	**.0002**	**.0000**	**.0000**
	0.16	**.0116**	**.0007**	**.0001**	**.0000**	**.0000**
	0.17	**.0084**	**.0004**	**.0000**	**.0000**	**.0000**
	0.18	**.0060**	**.0002**	**.0000**	**.0000**	**.0000**
	0.19	**.0041**	**.0001**	**.0000**	**.0000**	**.0000**
	0.2	**.0028**	**.0001**	**.0000**	**.0000**	**.0000**

<표 6> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과($N_1 = 500$)

오분류율 \ 표본크기		$N_1 = 500$				
		N_2				
		10%	20%	30%	40%	50%
0.05	0.01	.4136	.2687	.1934	.1481	.1184
	0.02	.3408	.1975	.1297	.0920	.0689
	0.03	.2764	.1410	.0835	.0543	.0378
	0.04	.2204	.0975	.0514	.0303	.0194
	0.05	.1725	.0651	.0302	.0160	.0093
	0.06	.1323	.0420	.0169	.0079	.0042
	0.07	.0993	.0260	.0089	.0037	.0017
	0.08	.0726	.0154	.0044	.0016	.0007
	0.09	.0517	.0087	.0021	.0006	.0002
	0.1	.0357	.0046	.0009	.0002	.0001
	0.11	.0238	.0023	.0004	.0001	.0000
	0.12	.0152	.0011	.0001	.0000	.0000
	0.13	.0093	.0005	.0000	.0000	.0000
	0.14	.0053	.0002	.0000	.0000	.0000
	0.15	.0029	.0001	.0000	.0000	.0000
	0.16	.0028	.0001	.0000	.0000	.0000
	0.17	.0016	.0000	.0000	.0000	.0000
	0.18	.0009	.0000	.0000	.0000	.0000
	0.19	.0005	.0000	.0000	.0000	.0000
	0.2	.0002	.0000	.0000	.0000	.0000

<표 7> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과($N_1 = 1000, \alpha = 1$)

오분류율 \ 표본크기		$N_1 = 1000$				
		N_2				
		10%	20%	30%	40%	50%
0.01	0.01	.7028	.6054	.5433	.4988	.4651
	0.02	.5667	.4380	.3615	.3100	.2729
	0.03	.4441	.3002	.2229	.1752	.1432
	0.04	.3373	.1941	.1267	.0894	.0666
	0.05	.2476	.1178	.0659	.0408	.0272
	0.06	.1752	.0667	.0311	.0166	.0097
	0.07	.1189	.0351	.0133	.0059	.0030
	0.08	.0772	.0170	.0050	.0018	.0008
	0.09	.0476	.0075	.0017	.0005	.0002
	0.1	.0277	.0030	.0005	.0001	.0000
	0.11	.0152	.0011	.0001	.0000	.0000
	0.12	.0077	.0003	.0000	.0000	.0000
	0.13	.0036	.0001	.0000	.0000	.0000
	0.14	.0015	.0000	.0000	.0000	.0000
	0.15	.0006	.0000	.0000	.0000	.0000
	0.16	.0004	.0000	.0000	.0000	.0000
	0.17	.0002	.0000	.0000	.0000	.0000
	0.18	.0001	.0000	.0000	.0000	.0000
	0.19	.0000	.0000	.0000	.0000	.0000
	0.2	.0000	.0000	.0000	.0000	.0000

<표 8> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과($N_1 = 1000, \alpha = 2$)

표본크기 / 오분류율		$N_1 = 1000$				
		N_2				
		10%	20%	30%	40%	50%
0.01	0.01	.6919	.5919	.5284	.4832	.4490
	0.02	.5390	.4062	.3290	.2779	.2416
	0.03	.4004	.2560	.1823	.1384	.1101
	0.04	.2814	.1461	.0881	.0584	**.0413**
	0.05	.1851	.0741	**.0363**	**.0203**	**.0124**
	0.06	.1124	**.0327**	**.0124**	**.0056**	**.0029**
	0.07	.0620	**.0122**	**.0034**	**.0012**	**.0005**
	0.08	**.0303**	**.0037**	**.0007**	**.0002**	**.0001**
	0.09	**.0127**	**.0009**	**.0001**	**.0000**	**.0000**
	0.1	**.0044**	**.0002**	**.0000**	**.0000**	**.0000**
	0.11	**.0012**	**.0000**	**.0000**	**.0000**	**.0000**
	0.12	**.0002**	**.0000**	**.0000**	**.0000**	**.0000**
	0.13	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.14	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.15	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.16	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.17	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.18	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.19	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.2	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**

<표 9> 삼차원 포화모형에서 w_{12}항의 유의성검정 결과($N_1 = 1000, \alpha = 3$)

표본크기 / 오분류율		$N_1 = 1000$				
		N_2				
		10%	20%	30%	40%	50%
0.01	0.01	.6759	.5718	.5064	.4601	.4253
	0.02	.5005	.3630	.2855	.2357	.2010
	0.03	.3438	.2020	.1348	.0972	.0741
	0.04	.2147	.0952	.0510	**.0307**	**.0201**
	0.05	.1185	**.0363**	**.0146**	**.0069**	**.0038**
	0.06	.0555	**.0105**	**.0029**	**.0010**	**.0004**
	0.07	**.0208**	**.0021**	**.0004**	**.0001**	**.0000**
	0.08	**.0057**	**.0003**	**.0000**	**.0000**	**.0000**
	0.09	**.0010**	**.0000**	**.0000**	**.0000**	**.0000**
	0.1	**.0001**	**.0000**	**.0000**	**.0000**	**.0000**
	0.11	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.12	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.13	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.14	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.15	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	.16	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.17	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.18	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.19	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**
	0.2	**.0000**	**.0000**	**.0000**	**.0000**	**.0000**

<표 10> 모형의 적합도검정 결과(N_1=500)

N_1	N_2	A	B	$C-\alpha$	$F-\alpha$	w	w_1	w_2	G^2	p-값
500	50	0.03	0.03	1	1.61983	-2.30259	0	0	0.6589	0.41696
			0.04	1	1.75758	-2.30238	0.02033	-0.002588	0.8997	0.34287
			0.05	1	1.90909	-2.30177	0.04089	-0.005954	1.1812	0.27712
			0.06	1	2.07656	-2.30073	0.06173	-0.010124	1.5057	0.21979
			0.07	1	2.26263	-2.29926	0.08294	-0.015138	1.8763	0.17075
			0.08	1	2.47059	-2.29735	0.10458	-0.021044	2.2967	0.12965
			0.09	1	2.70455	-2.29496	0.12675	-0.027898	2.7716	0.09595
			0.10	1	2.96970	-2.29208	0.14954	-0.035776	3.3065	0.06901
500	100	0.03	0.03	1	1.61983	-1.60944	0.00000	0.00000	1.2073	0.27186
			0.04	1	1.75758	-1.60924	0.02028	-0.002366	1.6482	0.19920
			0.05	1	1.90909	-1.60862	0.04075	-0.005438	2.1632	0.14135
			0.06	1	2.07656	-1.60759	0.06146	-0.009236	2.7565	0.09686
			0.07	1	2.26263	-1.60614	0.08248	-0.013792	3.4331	0.06390
			0.08	1	2.47059	-1.60424	0.10388	-0.019142	4.1996	0.04043*
			0.09	1	2.70455	-1.60188	0.12571	-0.025332	5.0637	0.02443*
			0.10	1	2.96970	-1.59904	0.14807	-0.032416	6.0353	0.01402*
500	150	0.03	0.03	1	1.61983	-1.20397	0.00000	0.00000	1.6710	0.19612
			0.04	1	1.75758	-1.20377	0.02024	-0.002180	2.2808	0.13098
			0.05	1	1.90909	-1.20316	0.04064	-0.005006	2.9927	0.08364
			0.06	1	2.07656	-1.20213	0.06125	-0.008494	3.8122	0.05088
			0.07	1	2.26263	-1.20068	0.08213	-0.012672	4.7459	0.02937*
			0.08	1	2.47059	-1.19880	0.10333	-0.017566	5.8023	0.01601*
			0.09	1	2.70455	-1.19646	0.12492	-0.023212	6.9917	0.00819*
			0.10	1	2.96970	-1.19365	0.14695	-0.029654	8.3267	0.00391*
500	200	0.03	0.03	1	1.61983	-0.91629	0.00000	0.00000	2.0682	0.15040
			0.04	1	1.75758	-0.91609	0.02020	-0.002020	2.8225	0.09295
			0.05	1	1.90909	-0.91548	0.04055	-0.004638	3.7028	0.05432
			0.06	1	2.07656	-0.91446	0.06109	-0.007864	4.7153	0.02990*
			0.07	1	2.26263	-0.91301	0.08185	-0.011722	5.8681	0.01542*
			0.08	1	2.47059	-0.91114	0.10291	-0.016234	7.1711	0.00741*
			0.09	1	2.70455	-0.90882	0.12429	-0.021428	8.6364	0.00330*
			0.10	1	2.96970	-0.90603	0.14608	-0.027340	10.279	0.00135*
500	250	0.03	0.03	1	1.61983	-0.69315	0.00000	0.00000	2.4122	0.12039
			0.04	1	1.75758	-0.69295	0.02018	-0.001884	3.2917	0.06963
			0.05	1	1.90909	-0.69234	0.04048	-0.004320	4.3174	0.03772*
			0.06	1	2.07656	-0.69123	0.06095	-0.007324	5.4967	0.01905*
			0.07	1	2.26263	-0.68988	0.08163	-0.010908	6.8385	0.00892*
			0.08	1	2.47059	-0.68801	0.10256	-0.015094	8.3538	0.00385*
			0.09	1	2.70455	-0.68570	0.12380	-0.019906	10.056	0.00152*
			0.10	1	2.96970	-0.68294	0.14538	-0.025370	11.963	0.00054*

참 고 문 헌

[1] 고 봉성 (1995). 범주형 자료의 오분류에 관한 연구, 성균관대학교 대학원 박사학위 논문.

[2] Agristi, A. (1990). *Categorical Data Analysis*, John Wiley & Sons.

[3] Anderson, E. R. (1991). *The Statistical Analysis of Categorical Data*, Springer-Verlag.

[4] Barron, B. A. (1977). The Effects of Misclassification on the Estimation of Relative Risk, *Biometrics*,. 33. 414-418.

[5] Bross, I. (1954). Misclassification in 2×2 Tables, *Biometrics*, 10, 478-486.

[6] Chen, T. Timothy. (1979). Log-Liner Models for Categorical Data with Misclassification and Double Sampling, *Journal of the American Statistical Association*, 74, 481- 488.

[7] Chen, T. and Fienberg, Stephen E. (1976). The Analysis of Contingency Tables with Incompletely Classified Data, *Biometrics*. 32, 133-144.

[8] Chen, T. (1974). Two-Dimensional Contingency Tables with Both Completely and Partially Cross-Classified Data, *Biometrics*, 30, 629-642.

[9] Chiacchierini, R. P., and Arnold. J. C. (1977). A two sample test for independence in 2×2 contingency tables with both margins subjects to misclassification, *Journal of the American Statistical Association*, 72, 170-174.

[10] Christensen, R. (1990). *Log-Linear Models*, Springer-Verlag.

[11] Collet, D. (1991). *Modelling Binary Data,* Chapman and Hall, London.

[12] Fienberg, S. E. (1981). *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge, Mass.

[13] Goldberg, T. D. (1975). The Effects of Misclassification on the Bias in the Difference Between Two Proportions and the Relative Odds in the Fourfold Table, *Journal of the American Statistical Association*, 70, 561-567.

[14] Hochberg, Y. (1977). On the use of double sampling schemes in analysing categorical data with misclassification error, *Journal of the American Statistical Association*, 72, 914-921.

[15] Korn, E. L. (1981). Hierarchical Log-linear Models not Preserved by Classification error, *Journal of the American Statistical Association*, 76, 110-113.

[16] McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

[17] Rogot, E. (1961). A Note on Measurement Errors and Detecting Real Differences, *Journal of the American Statistical Association*, 56, 314-319.

[18] Tenenbein, A. (1970). A Double Sampling Scheme for Estimating from Binomial Data with Misclassification, *Journal of the American Statistical Association*, 65, 1350-1361.

[19] Tenenbein, A. (1972). A Double Sampling Scheme for Estimating From Misclassified Multinomial Data with Application to Sampling Inspection, *Technometrics*, 14, 187-202.