한국통계학회논문집
2010, 17권, 1호, 67-77

단계별로 얻어진 이차원 분할표의 모수 추정을 위한 정확최대우도추정법과 단계별추출추정법의 비교

이상은[a], 강기훈[b], 정석오[b], 신기일[1,b]

[a]경기대학교 응용정보통계학과, [b]한국외국어대학교 통계학과

요 약

단계별로 얻어진 $I \times J$ 이차원 범주형 자료에서 분할표 일부의 칸에서 도수가 붕괴(collapse)된 상태로 조사가 이루어진 것을 단계별추출(step-wise sampling)이라 한다. 단계별추출로 얻어진 자료를 분석할 경우 단계별추출법을 사용하여 분석하면 분석의 효과를 얻을 수 있다. 본 논문에서는 단계별추출법 중에서 최대우도추정법을 이용하여 얻어진 정확최대우도추정량(exact maximum likelihood estimator)과 단계별추출최대우도추정량을 연구하였다. 또한 MSE와 편향(bias)을 기준으로 모의실험을 통하여 두 추정법을 비교하였다.

주요용어: 단계별추출추정법, 최대우도추정법, 분할표, Newton-Raphson 방법.

1. 서론

두 범주가 독립이고 I개의 수준과 J개의 수준을 갖고 있는 이차원 분할표의 각 칸이 단계별로 얻어졌다고 가정하자. 이때 각 단계에서 얻어진 칸별 도수가 모든 칸에서 얻어지지 않고 어떤 단계에서 특정 칸이 붕괴(collapse) 되었다고 하자. 예를 들어 2×2 이차원 분할표에서 자료가 얻어졌다고 가정하자. 자료는 단계별로 얻어지며 최종 3 단계까지 얻어진다. 1 단계에서는 (1,1)칸의 자료가 얻어지고 다른 칸은 붕괴되어 하나의 자료만 얻어진다. 2 단계는 (1,1)칸 자료와 (2,2)칸 자료 그리고 다른 두 칸은 붕괴되어 얻어진다. 최종 3 단계는 모든 칸, 즉 (1,1), (1,2), (2,1) 그리고 (2,2) 칸의 자료가 얻어진다. 이렇게 얻어진 자료가 단계별추출법으로 얻어진 자료이다. 이런 경우 붕괴된 자료를 사용하지 않고 완전히 얻어진 3 단계 자료만을 사용한다면 모든 단계에서 얻어진 자료를 사용하는 것에 비해 그 효율이 떨어지게 될 것이다. 이와 같이 일정 단계에서 붕괴된 자료가 얻어졌을 때 일반적으로 고려할 수 있는 방법이 단계별추출추정법(step-wise sampling estimation)이다. 1차원 분할표 결과이긴 하지만 Lee와 Park (1999)은 단계별추출추정법이 완전한 자료만을 사용했을 경우에 비해 매우 우수한 결과를 주는 것을 확인하였다.

Blumenthal (1968)은 범주형 자료 분석의 일부 자료가 얻어진 경우에서의 최대우도추정량에 관하여 연구하였으며 Hocking과 Oxspring (1971) 또한 일부 자료가 얻어진 범주형 자료에 관하여 연구하였다. 국내에서도 Lee와 Park (1999)은 Blumenthal (1968)과 Hocking과 Oxspring (1971) 연구의 특별한 경우인 단계별추출법에 관한 연구를 하였다. 또한 Park 등 (2006)의 연구는 최대우도추정량과 베이즈 추정량에 관한 연구를 하였다. 일반적으로 단계별추출추정법에 사용되는 방법은 최대우도추정법이다. 현재 1차원 분할표를 위한 단계별추출최대우도추정법에 관한 연구 결과가 Lee와 Park (1999) 그

이 연구는 2009년 한국외국어대학교 교내연구비의 지원에의해 이루어진 것임.
[1] 교신저자: (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과, 교수.
E-mail: keyshin@hufs.ac.kr

리고 Park 등 (2006)에 나와 있다. 그러나 많은 자료가 2차원 분할표로 이루어져 있기 때문에 2차원 분할표에 관한 단계별추출최대우도추정법에 관한 연구가 절실히 필요하다.

먼저 단계별추출추정법을 살펴보기 위해 서로 독립인 두 범주의 2×2 분할표를 고려하자. 두 범주가 독립이므로 각 칸의 확률은 $(p_i \times q_j)$, $i = 1, 2$, $j = 1, 2$로 계산될 수 있다. 여기서 p_1은 첫 번째 범주의 성공확률이고 q_1은 두 번째 범주의 성공확률이다. 따라서 $p_1 \times q_1$은 두 범주 모두 성공인 확률을 의미한다. 또한 $p_1 + p_2 = 1$이고 $q_1 + q_2 = 1$이다. 다음으로 각 칸에서 얻어진 도수를 $f_{ij}^{(k)}$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3$이라 하자. 여기서 i는 첫 번째 범주를 j는 두 번째 범주를 그리고 k는 단계를 나타내며 본 논문에서는 3단계가 있다고 가정하였다. 이제 각 단계별로 얻어진 자료의 붕괴상태를 살펴보자. 1단계는 $f_{11}^{(1)}$과 $f_{12}^{(1)} + f_{21}^{(1)} + f_{22}^{(1)}$이 얻어졌다. 즉 $f_{11}^{(1)}$을 제외한 다른 칸이 붕괴된 상태로 자료가 얻어졌다. 2단계는 $f_{11}^{(2)}$, $f_{22}^{(2)}$ 그리고 $f_{12}^{(2)} + f_{21}^{(2)}$가 얻어졌다. 즉 $f_{12}^{(2)}$와 $f_{21}^{(2)}$가 붕괴된 것이다. 다음으로 3단계에서는 붕괴가 일어나지 않아 완전한 자료가 얻어졌다고 가정하자. 즉 $f_{11}^{(3)}$, $f_{12}^{(3)}$, $f_{21}^{(3)}$, $f_{22}^{(3)}$가 모두 얻어졌다. 일반적으로 완전한 자료인 3단계만을 이용하여 분석할 경우 쉽게 모수를 추정할 수 있을 것이다. 그러나 이는 1단계와 2단계가 갖고 있는 정보를 이용하지 않았기 때문에 정보의 손실을 가져오게 되어 모든 정보를 사용하여 얻은 분석에 비해 추정의 정도가 나빠지게 된다. 본 논문에서는 1 단계에서 3 단계의 자료를 모두 사용하여 모수 p_i와 q_j를 추정하는 방법인 단계별추출 추정법을 연구하였다. 이 연구는 $I \times J$ 분할표인 경우에도 적용이 가능하다. 그러나 여러 경우가 발생할 수 있고 또한 수식이 복잡하여 본 연구에서는 2×2인 경우만을 살펴보았다.

본 논문에서는 각 단계별 우도함수를 구한 후 이를 이용하여 결합우도함수를 구하였다. 그러나 얻어진 우도함수를 최대로 하는 최대우도추정량 공식(closed form)을 구하는 것은 쉽지가 않다. 이에 본 논문에서는 정확최대우도추정량을 구하기 위해 수치해석적인 방법을 사용하여 추정치를 구하였다. 다음으로 정확최대우도추정량의 대안으로 단계별추출최대우도추정량을 제안하였다. 이 방법은 이미 제안된 1차원 분할표를 위한 단계별추출최대우도추정량을 확장한 것이며 얻어진 2차원 분할표를 위한 단계별추출최대우도추정량은 통계량의 형태가 매우 간단하게 구해진다.

본 논문은 먼저 2장에서 각 단계별 우도함수를 계산한 후 이를 결합한 결합우도함수를 유도하였다. 3장에서는 모수변환(re-parametrization)을 이용한 단계별최대우도추정량을 구하였다. 4장에 모의실험 결과가 있으며 결론은 5장에 있다.

2. 정확최대우도추정량을 위한 단계별 우도함수

이 장에서는 각 단계에서 얻어진 정보를 모두 합쳐 최종 결합우도함수를 구한 후 이를 기반으로 정확최대우도추정량을 구하였다. 최종 결합우도함수를 구하기 위해 먼저 각 단계별 우도함수를 구했다. 다음이 우도함수를 구하기 위해 얻어진 각 단계별 도수다.

1 단계 표본 도수 : $f_{11}^{(1)}$, $n_1 - f_{11}^{(1)}$

2 단계 표본 도수 : $f_{11}^{(2)}$, $f_{22}^{(2)}$, $n_2 - f_{11}^{(2)} - f_{22}^{(2)}$

3 단계 표본 도수 : $f_{11}^{(3)}$, $f_{22}^{(3)}$, $f_{21}^{(3)}$, $f_{12}^{(3)} = n_3 - f_{11}^{(3)} - f_{22}^{(3)} - f_{21}^{(3)}$,

여기서 $f_{ij}^{(k)}$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3$는 각 칸에서 얻어진 도수를 나타내며 i는 첫 번째 범주를 j는 두 번째 범주를 그리고 k는 단계를 나타낸다. 또한 n_1, n_2 그리고 n_3는 각 단계별로 얻어진 표본 크기이다. 다음으로 구해진 각 단계별 우도함수는 다음과 같다. 먼저 $p_1 = p$, $p_2 = 1 - p$, $q_1 = q$ 그리고 $q_2 = 1 - q$라 하자. 그러면 1 단계의 자료를 이용한 우도함수는 다음과 같다.

$$L(p, q|\text{data}) \propto (pq)^{f_{11}^{(1)}} (1 - pq)^{n_1 - f_{11}^{(1)}}. \tag{2.1}$$

이때 헤시안 행렬(Hessian matrix)은 다음과 같다.

$$H_1 = \frac{n_1}{1-pq}\begin{pmatrix} q/p & 1 \\ 1 & p/q \end{pmatrix}.$$

다음으로 2 단계에서 얻은 우도함수는 다음과 같으며

$$L(p,q|\text{data}) \propto (pq)^{f_{11}^{(2)}}\{(1-p)(1-q)\}^{f_{22}^{(2)}} \times \{1-pq-(1-p)(1-q)\}^{n_2-f_{11}^{(2)}-f_{22}^{(2)}} \tag{2.2}$$

또한 헤시안 행렬(Hessian matrix)은 다음과 같다.

$$H_2 = n_2\begin{pmatrix} H_{2(11)} & H_{2(12)} \\ H_{2(21)} & H_{2(22)} \end{pmatrix},$$

여기서 $H_{2(11)} = (p+q-2pq)/\{p(1-p)\} + (1-2q)2/(p+q-2pq), H_{2(12)} = 1/(p+q-2pq), H_{2(21)} = 1/(p+q-2pq)$이고 $H_{2(22)} = (p+q-2pq)/\{q(1-q)\} + (1-2p)2/(p+q-2pq)$이다.

다음으로 3 단계에서 얻은 우도함수는

$$L(p,q|\text{data}) \propto (pq)^{f_{11}^{(3)}}\{(1-p)(1-q)\}^{f_{22}^{(3)}}\{(1-p)q\}^{n_3-f_{11}^{(3)}-f_{22}^{(3)}-f_{21}^{(3)}} \times \{p(1-q)\}^{f_{21}^{(3)}} \tag{2.3}$$

으로 구해지며 헤시안 행렬(Hessian matrix)은 다음과 같다.

$$H_3 = n_3\begin{pmatrix} \dfrac{1}{p(1-p)} & 0 \\ 0 & \dfrac{1}{q(1-q)} \end{pmatrix}.$$

최종적으로 1 단계에서 3 단계를 모두 합쳐 구한 최종 결합우도함수는 다음과 같으며

$$\begin{aligned} L(p,q|\text{data}) \propto\ & (pq)^{f_{11}^{(1)}+f_{11}^{(2)}+f_{11}^{(3)}}\{(1-p)(1-q)\}^{f_{22}^{(2)}+f_{22}^{(3)}}\{(1-p)q\}^{n_3-f_{11}^{(3)}-f_{22}^{(3)}-f_{21}^{(3)}} \\ & \times \{p(1-q)\}^{f_{21}^{(3)}}\{1-(pq)\}^{n_1-f_{11}^{(1)}} \times \{1-pq-(1-p)(1-q)\}^{n_2-f_{11}^{(2)}-f_{22}^{(2)}}. \end{aligned} \tag{2.4}$$

모든 단계을 결합하여 얻은 헤시안 행렬을 $H^{(3)}$이라 하면 $H^{(3)} = H_3 + H_2 + H_1 = H_3 + H^{(2)}$이 성립하고 이는 다음과 같이 표시될 수 있다.

$$\left(H^{(3)}\right)^{-1} = H_3^{-1} - H_3^{-1}H^{(2)}\left(H^{(3)}\right)^{-1}.$$

따라서 모든 정보를 이용하여 얻은 추정량 $\hat{p}$과 $\hat{q}$의 분산은 3 단계만을 사용하여 구한 분산에 비해 작아지게 된다. 물론 이 결론이 만족되기 위해서는 행렬

$$H_3^{-1}H^{(2)}\left(H^{(3)}\right)^{-1}$$

이 양정치 또는 비음정치 행렬이라는 조건이 필요하다. 이 행렬의 추정값은 p와 q 대신에 추정값을 대입함으로써 구할 수 있다.

최대우도추정량을 얻기 위해서는 식 (2.4)를 미분하여 방정식을 풀어야 한다. 그러나 그 결과가 복잡하기 때문에 본 논문에서는 수치해석적인 방법으로 이를 추정하였다. 수치해석에서 일반적으로 사용하는 방법은 Newton-Raphson 방법이며 본 논문에서 사용한 Newton-Raphson의 반복적 방법은 다음의 수식을 이용하였다. 자료분석에서는 SAS의 Proc/nlin이 사용되었다.

$$\begin{pmatrix} p(i+1) \\ q(i+1) \end{pmatrix} = \begin{pmatrix} p(i) \\ q(i) \end{pmatrix} - \left(H^{(3)}\right)^{-1}\frac{d}{d(p,q)}\ln L(p(i),q(i)|\text{data}).$$

3. 단계별 우도함수를 이용한 단계별추출최대우도추정량

2장에서 설명한 바와같이 정확최대우도추정치는 일반적으로 수치해석적으로 구해진다. 본 논문은 Park 등 (2006)에서 소개한 단계별추출추정방법을 적용함으로써 단계별추출최대우도추정량공식(closed form)을 제시하였다.

단계별추출최대우도추정량을 구하기 위하여 다음과 같이 모수변환(reparameterization)을 실시하였다. 먼저 2장에서와 같이 $p_1 = p$, $p_2 = 1 - p$, $q_1 = q$ 그리고 $q_2 = 1 - q$라 하자. 그러면 변환된 모수는 다음과 같다.

$$\theta_1 = pq, \qquad \theta_2 = (1-p)(1-q), \qquad \theta_3 = p(1-q), \qquad \theta_4 = (1-p)q.$$

또한 변환된 단계별 모수와 이때 얻어진 도수는 다음과 같다.

단계	모수	표본 도수
1.	$\theta_1, 1-\theta_1$	$f_{11}^{(1)}, n_1 - f_{11}^{(1)}$
2.	$\theta_1, \theta_2, 1-\theta_1-\theta_2$	$f_{11}^{(2)}, f_{22}^{(2)}, n_2 - f_{11}^{(2)} - f_{22}^{(2)}$
3.	$\theta_1, \theta_2, \theta_3, \theta_4 = 1-\theta_1-\theta_2-\theta_3$	$f_{11}^{(3)}, f_{22}^{(3)}, f_{21}^{(3)}, f_{12}^{(3)} = n_3 - f_{11}^{(3)} - f_{22}^{(3)} - f_{21}^{(3)}$

다음으로 각 단계 만을 사용하여 얻어진 최대우도추정량을 살펴보면 다음과 같다. 이에 관한 내용은 Park 등 (2006)을 참고하기 바란다.

먼저 1 단계 만을 이용하여 얻은 최대우도추정량 결과는 다음과 같다.

$$\widehat{\theta_1} = \frac{f_{11}^{(1)}}{n_1},$$

$$\widehat{(1-\theta_1)} = \frac{n_1 - f_{11}^{(1)}}{n_1}.$$

다음으로 2 단계 만을 이용하여 얻은 최대우도추정량은

$$\widehat{\theta_1} = \frac{f_{11}^{(2)}}{n_2}, \qquad \widehat{\theta_2} = \frac{f_{22}^{(2)}}{n_2},$$

$$\widehat{(1-\theta_1-\theta_2)} = \frac{n_2 - f_{11}^{(2)} - f_{22}^{(2)}}{n_2}$$

이며 3 단계 만을 이용하여 얻은 최대우도추정량 결과는 다음과 같다.

$$\widehat{\theta_1} = \frac{f_{11}^{(3)}}{n_3}, \qquad \widehat{\theta_2} = \frac{f_{22}^{(3)}}{n_3}, \qquad \widehat{\theta_3} = \frac{f_{12}^{(3)}}{n_3},$$

$$\widehat{(1-\theta_1-\theta_2-\theta_3)} = \frac{n_3 - f_{11}^{(3)} + f_{12}^{(3)} - f_{12}^{(3)}}{n_3}.$$

이제 1 단계와 2 단계를 이용하여 얻은 결합우도함수를 이용하여 최대우도추정량을 구하면 다음과 같으며

$$(n_1+n_2)(\hat{\theta}_1) = (f_{11}^{(1)} + f_{11}^{(2)}),$$

$$(n_1+n_2)(\hat{\theta}_2) = f_{22}^{(2)} + (n_1 - f_{11}^{(1)})\left(\frac{f_{22}^{(2)}}{n_2 - f_{11}^{(2)}}\right).$$

끝으로 모든 단계를 결합하여 얻어진 결합우도함수를 이용하여 얻어진 최대우도추정량은 다음의 식으로 얻을 수 있다.

$$
\begin{aligned}
(n_1+n_2+n_3)\hat{\theta_{1M}} &= f_{11}^{(1)} + f_{11}^{(2)} + f_{11}^{(3)},\\
(n_1+n_2+n_3)\hat{\theta_{2M}} &= f_{22}^{(2)} + f_{22}^{(3)} + \frac{\left(n_1 - f_{11}^{(1)}\right)\left(f_{22}^{(2)} + f_{22}^{(3)}\right)}{n_2 + n_3 - f_{11}^{(2)} - f_{11}^{(3)}},\\
(n_1+n_2+n_3)\hat{\theta_{3M}} &= f_{21}^{(3)} + \left(n_2 - f_{11}^{(2)} - f_{22}^{(2)}\right)\left(\frac{f_{21}^{(3)}}{n_3 - f_{11}^{(3)} - f_{22}^{(3)}}\right) + \left(n_1 - f_{11}^{(1)}\right)\\
&\quad \times\left[f_{21}^{(3)} + \left(n_2 - f_{11}^{(2)} - f_{22}^{(2)}\right)\left(\frac{f_{21}^{(3)}}{n_3 - f_{11}^{(3)} - f_{22}^{(3)}}\right)\right] \Big/ \left(n_2 - f_{11}^{(2)} + n_3 - f_{11}^{(3)}\right),\\
\hat{\theta_{4M}} &= 1 - \hat{\theta_{1M}} - \hat{\theta_{2M}} - \hat{\theta_{3M}}.
\end{aligned}
$$

이제 변환된 모수를 재변환하면 최종적으로 모수 p와 q의 추정식을 다음과 같이 얻을 수 있다.

$$
\begin{aligned}
\hat{p_M} &= \hat{\theta_{1M}} + \hat{\theta_{3M}},\\
\hat{q_M} &= \hat{\theta_{1M}} + \hat{\theta_{4M}} \quad \text{or} \quad \widehat{(1-q)} = \hat{\theta_{2M}} + \hat{\theta_{3M}}.
\end{aligned}
$$

4. 모의실험

모의실험은 다음과 같이 실시하였다. 먼저 자료 수 n은 $n = 20, 50, 150$ 그리고 200을 사용하였다. 또한 모수 p, q의 참값으로 $p \neq q$인 경우에는 (p, q) = (0.1, 0.9), (0.3, 0.7), (0.5, 0.5), (0.7, 0.3) 그리고 (0.9, 0.1)을 사용하였다. 이 경우를 case 1이라 하였다. 또한 모수 참값이 같은 $p = q$ 경우에는 (p, q) = (0.1, 0.1), (0.3, 0.3), (0.5, 0.5), (0.7, 0.7) 그리고 (0.9, 0.9)를 사용하여 자료를 생성하였다. 이 경우는 case 2라 하였다. 모의실험은 5,000번 반복하여 실시하였고 사용된 비교 통계량은 MSE와 편향, Bias이며 정의는 다음과 같다.

$$
\begin{aligned}
\text{MSE} &= \frac{1}{5000}\sum_{r=1}^{5000}(\hat{p} - p)^2,\\
\text{Bias} &= \frac{1}{5000}\sum_{r=1}^{5000}(\hat{p} - p).
\end{aligned}
$$

정확최대우도추정량의 경우에는 Newton-Raphson 방법이 사용되었으며 본 모의실험에서는 SAS의 Proc/nlin을 사용하여 모수를 추정하였다. 또한 단계별추출최대우도추정량은 3장의 결과를 사용하였다. 모의실험 결과는 표 1~3에 정리하였다. 그림 1과 2는 그림의 모양을 위해 표 3에서 얻어진 값의 절대값(절대편향)을 이용하였다.

표 1은 모수 p, q의 추정값을 나타내며 표 2는 MSE를 그리고 표 3은 편향 결과를 나타낸다. 먼저 추정값의 결과인 표 1을 살펴보면 자료의 수와 모수 값에 상관없이 두 방법의 추정값이 모두 유사하며 그 값은 참값에 매우 근사하다. 다음으로 MSE의 결과인 표 2를 살펴보면 정확최대우도추정량이 예상대로 매우 우수한 결과를 주고 있으며 두 방법의 결과를 비교하면 정확최대우도추정량의 MSE가 우수하다. 그러나 단계별최대우도추정량 또한 우수한 결과를 주고 있다. 이러한 패턴은 자료의 수와 모수 p, q값에 상관없이 일정하다. 다음으로 편향 결과인 표 3을 살펴보자. 표 3의 결과에서는 단계별추출

표 1: 정확최대우도추정량과 단계별최대우도추정량 비교

$n_i = n$	참값		MLE_{ex}		MLE_{step}	
	p	q	$\hat{p}_{M(ex)}$	$\hat{q}_{M(ex)}$	$\hat{p}_{M(step)}$	$\hat{q}_{M(step)}$
20	0.1	0.9	0.111413675	0.894989056	0.099169939	0.899673138
	0.3	0.7	0.308993483	0.697698368	0.301557334	0.700326149
	0.5	0.5	0.500903330	0.502710974	0.499500613	0.501691757
	0.7	0.3	0.701733419	0.305620001	0.699860490	0.300240382
	0.9	0.1	0.888320150	0.110629183	0.900593960	0.099711473
	0.1	0.1	0.114573699	0.113808146	0.101312413	0.100826066
	0.3	0.3	0.301095225	0.312254106	0.300316650	0.299291342
	0.5	0.5	0.499337195	0.503708262	0.502529117	0.498904903
	0.7	0.7	0.695297560	0.696401528	0.699838129	0.699863811
	0.9	0.9	0.892852704	0.892413266	0.899828838	0.899328923
50	0.1	0.9	0.105051895	0.895669191	0.099824561	0.899749065
	0.3	0.7	0.301872406	0.699763491	0.299572232	0.699006046
	0.5	0.5	0.500709671	0.499244133	0.500246606	0.499772670
	0.7	0.3	0.698653780	0.301178333	0.698867496	0.300476898
	0.9	0.1	0.897712972	0.100217990	0.899979545	0.099942423
	0.1	0.1	0.105691423	0.102775155	0.099889378	0.099834714
	0.3	0.3	0.304466673	0.301845980	0.300113198	0.299639894
	0.5	0.5	0.501431720	0.500313597	0.500046175	0.500106470
	0.7	0.7	0.698033886	0.697586365	0.700720362	0.699427940
	0.9	0.9	0.895513635	0.895131099	0.900435307	0.900294548
150	0.1	0.9	0.100034332	0.899663063	0.100035659	0.899955569
	0.3	0.7	0.299867455	0.698478318	0.300053727	0.699914704
	0.5	0.5	0.502366908	0.498839651	0.501095764	0.499682508
	0.7	0.3	0.701228802	0.300201652	0.700002373	0.299755201
	0.9	0.1	0.899405774	0.100745049	0.899950480	0.100056578
	0.1	0.1	0.100587378	0.100641408	0.099870614	0.099505756
	0.3	0.3	0.300717809	0.300269501	0.300481520	0.300036653
	0.5	0.5	0.501168739	0.498679865	0.500380508	0.500132660
	0.7	0.7	0.700560398	0.700040257	0.700296302	0.700008878
	0.9	0.9	0.897810248	0.899942491	0.900094948	0.899705145
200	0.1	0.9	0.100544985	0.899406149	0.099876142	0.899727259
	0.3	0.7	0.300526947	0.699697080	0.299718533	0.700166387
	0.5	0.5	0.500781919	0.498636066	0.499862929	0.500078978
	0.7	0.3	0.700869272	0.299024540	0.700197882	0.299481254
	0.9	0.1	0.900016163	0.100421864	0.900177935	0.099927439
	0.1	0.1	0.100499813	0.101153729	0.100403672	0.099760283
	0.3	0.3	0.298106095	0.301094492	0.299698429	0.300026408
	0.5	0.5	0.499648558	0.500335409	0.499920361	0.500522090
	0.7	0.7	0.699644606	0.700948389	0.699066010	0.700871407
	0.9	0.9	0.899685840	0.899655449	0.899932500	0.900519656

최대우도추정량이 정확최대우도추정량에 비해 편향이 작은 것을 확인 할 수 있다. 이를 쉽게 알 수 있도록 표 3의 결과를 절대편향을 이용하여 그림으로 나타냈다. 그림 1과 2를 살펴보면 $n = 20$인 경우 절대편향은 차이를 보이고 있으며 자료 수가 증가하면 절대편향의 차이는 예상대로 줄어드는 경향을 보이고 있다. 물론 값 자체는 크지 않기 때문에 실제 자료 분석에서는 큰 영향을 주지 않는다고 판단된다. 결론적으로 MSE기준으로는 정확최대우도추정량이 우수하고 편향을 기준으로 하였을 때는 단계별추출최대우도추정량이 우수한 것으로 나타났다.

표 2: 정확최대우도추정량과 단계별최대우도추정량의 MSE 비교

n / $\hat{n}$	참값		MLE_{ex}		MLE_{step}	
	p	q	$MSE_{p;ex}$	$MSE_{q;ex}$	$MSE_{p;step}$	$MSE_{q;step}$
20	0.1	0.9	0.000130272	2.51096E-05	0.001853600	0.002508103
	0.3	0.7	8.08827E-05	5.29751E-06	0.006306539	0.007052770
	0.5	0.5	8.16005E-07	7.34938E-06	0.008642149	0.009007059
	0.7	0.3	3.00474E-06	3.15844E-05	0.007102106	0.006295276
	0.9	0.1	0.000136419	0.000112900	0.002406997	0.001846592
	0.1	0.1	0.000212393	0.000190665	0.003712756	0.003616319
	0.3	0.3	1.19952E-06	0.000150163	0.007393211	0.007418725
	0.5	0.5	4.39310E-07	1.37512E-05	0.008935809	0.009160790
	0.7	0.7	2.21129E-05	1.29490E-05	0.007483879	0.007239110
	0.9	0.9	5.10838E-05	5.75585E-05	0.003564741	0.003446236
50	0.1	0.9	2.55216E-05	1.87559E-05	0.000764744	0.000975717
	0.3	0.7	3.50590E-06	5.59365E-08	0.002499399	0.002841294
	0.5	0.5	5.03633E-07	5.71335E-07	0.003485032	0.003594068
	0.7	0.3	1.81231E-06	1.38847E-06	0.002867101	0.002497180
	0.9	0.1	5.23050E-06	4.75198E-08	0.000974659	0.000713742
	0.1	0.1	3.23923E-05	7.70148E-06	0.001386135	0.001372032
	0.3	0.3	1.99512E-05	3.40764E-06	0.003080000	0.003016033
	0.5	0.5	2.04982E-06	9.83433E-08	0.003490318	0.003543073
	0.7	0.7	3.86561E-06	5.82563E-06	0.002931581	0.002784396
	0.9	0.9	2.01275E-05	2.37062E-05	0.001281181	0.001243695
150	0.1	0.9	1.17868E-09	1.13527E-07	0.000236843	0.000333132
	0.3	0.7	1.75681E-08	2.31552E-06	0.000799376	0.000928875
	0.5	0.5	5.60225E-06	1.34641E-06	0.001135178	0.001158100
	0.7	0.3	1.50995E-06	4.06633E-08	0.000963201	0.000822612
	0.9	0.1	3.53104E-07	5.55098E-07	0.000330600	0.000243206
	0.1	0.1	3.45013E-07	4.11404E-07	0.000449826	0.000420882
	0.3	0.3	5.15250E-07	7.26308E-08	0.001004654	0.000978460
	0.5	0.5	1.36595E-06	1.74276E-06	0.001160131	0.001102663
	0.7	0.7	3.14046E-07	1.62061E-09	0.000918684	0.000963041
	0.9	0.9	4.79501E-06	3.30724E-09	0.000415140	0.000402772
200	0.1	0.9	2.97009E-07	3.52659E-07	0.000178510	0.000247089
	0.3	0.7	2.77673E-07	9.17608E-08	0.000612531	0.000670004
	0.5	0.5	6.11398E-07	1.86032E-06	0.000854521	0.000868850
	0.7	0.3	7.55635E-07	9.51522E-07	0.000688482	0.000595456
	0.9	0.1	2.61236E-10	1.77969E-07	0.000246367	0.000181408
	0.1	0.1	2.49813E-07	1.33109E-06	0.000330857	0.000331715
	0.3	0.3	3.58688E-06	1.19791E-06	0.000781495	0.000746898
	0.5	0.5	1.23512E-07	1.12499E-07	0.000866157	0.000873037
	0.7	0.7	1.26305E-07	8.99441E-07	0.000721638	0.000738965
	0.9	0.9	9.86964E-08	1.18715E-07	0.000303879	0.000305452

5. 결론

단계별로 일부 범주형 자료의 도수가 붕괴된 경우에 사용하는 방법인 단계별추출추정법은 완전히 얻어진 자료만을 사용하여 분석한 경우에 비해 우수한 결과를 준다. 본 논문에서는 1차원 분할표의 확장인 2차원 분할표에서 사용할 수 있는 단계별추출추정법을 연구하였다. 각 단계를 모두 포함한 결합우도함수를 이용하여 정확최대우도함수를 추정할 수 있으나 이 경우에는 공식형태로 추정량이 얻어지

표 3: 정확최대우도추정량과 단계별최대우도추정량의 절대편향 비교

n	참값		MLE_{ex}		MLE_{step}	
	p	q	$Bias_{p:ex}$	$Bias_{q:ex}$	$Bias_{p:step}$	$Bias_{q:step}$
20	0.1	0.9	0.011413675	−0.005010944	−0.000830061	−0.000326862
	0.3	0.7	0.008993483	−0.002301632	0.001557334	0.000326149
	0.5	0.5	0.000903330	0.002710974	−0.000499387	0.001691757
	0.7	0.3	0.001733419	0.005620001	−0.000139510	0.000240382
	0.9	0.1	−0.011679850	0.010629183	0.000593960	−0.000288527
	0.1	0.1	0.014573699	0.013808146	0.001312413	0.000826066
	0.3	0.3	0.001095225	0.012254106	0.000316650	−0.000708658
	0.5	0.5	−0.000662805	0.003708262	0.002529117	−0.001095097
	0.7	0.7	−0.004702440	−0.003598472	0.000161871	−0.000136189
	0.9	0.9	−0.007147296	−0.007586734	0.000171162	−0.000671077
50	0.1	0.9	0.005051895	−0.004330809	−0.000175439	−0.000250935
	0.3	0.7	0.001872406	−0.000236509	−0.000427768	−0.000993954
	0.5	0.5	0.000709671	−0.000755867	0.000246606	−0.000227330
	0.7	0.3	−0.001346220	0.001178333	−0.001132504	0.000476898
	0.9	0.1	−0.002287028	0.000217990	−2.04553E−05	−5.75770E−05
	0.1	0.1	0.005691423	0.002775155	−0.000110622	−0.000165286
	0.3	0.3	0.004466673	0.001845980	0.000113198	−0.000360106
	0.5	0.5	0.001431720	0.000313597	4.61746E−05	0.000106470
	0.7	0.7	−0.001966114	−0.002413635	0.000720362	−0.000572060
	0.9	0.9	−0.004486365	−0.004868901	0.000435307	0.000294548
150	0.1	0.9	3.43319E−05	−0.000336937	3.56590E−05	−4.44306E−05
	0.3	0.7	−0.000132545	−0.001521682	5.37273E−05	−8.52956E−05
	0.5	0.5	0.002366908	−0.001160349	0.001095764	−0.000317492
	0.7	0.3	0.001228802	0.000201652	2.37323E−06	−0.000244799
	0.9	0.1	−0.000594226	0.000745049	−4.95205E−05	5.65780E−05
	0.1	0.1	0.000587378	0.000641408	−0.000129386	−0.000494244
	0.3	0.3	0.000717809	0.000269501	0.000481520	3.66532E−05
	0.5	0.5	0.001168739	−0.001320135	0.000380508	0.000132660
	0.7	0.7	0.000560398	4.02568E−05	0.000296302	8.87834E−06
	0.9	0.9	−0.002189752	−5.75086E−05	9.49476E−05	−0.000294855
200	0.1	0.9	0.000544985	−0.000593851	−0.000123858	−0.000272741
	0.3	0.7	0.000526947	−0.000302920	−0.000281467	0.000166387
	0.5	0.5	0.000781919	−0.001363934	−0.000137071	7.89783E−05
	0.7	0.3	0.000869272	−0.000975460	0.000197882	−0.000518746
	0.9	0.1	1.61628E−05	0.000421864	0.000177935	−7.25606E−05
	0.1	0.1	0.000499813	0.001153729	0.000403672	−0.000239717
	0.3	0.3	−0.001893905	0.001094492	−0.000301571	2.64084E−05
	0.5	0.5	−0.000351442	0.000335409	−7.96392E−05	0.000522090
	0.7	0.7	−0.000355394	0.000948389	−0.000933990	0.000871407
	0.9	0.9	−0.000314160	−0.000344551	−6.74999E−05	0.000519656

지 않기 때문에 사용에 제약이 있을 수 있다. 이에 반해 단계별추출최대우도추정량은 공식형태로 추정량이 주어지기 때문에 추정량의 통계적 성질 등을 규명하는데 강점이 있을 수 있다. 물론 표 2의 결과를 보면 MSE를 기준으로 하였을 때 정확최대우도추정량이 우수한 결과를 주고 있다. 그러나 단계별최대우도추정량의 경우에도 매우 작은 값의 MSE를 보여주고 있으며 표 3 그리고 그림 1과 2의 결과인 편향과 절대편향을 비교해 본 결과는 오히려 단계별추출최대우도추정량이 우수한 것으로 나타났다. 따라서 실제 자료 분석에서는 두 방법 중 어떤 방법을 사용해도 무방하다고 판단되나 추정을 위해

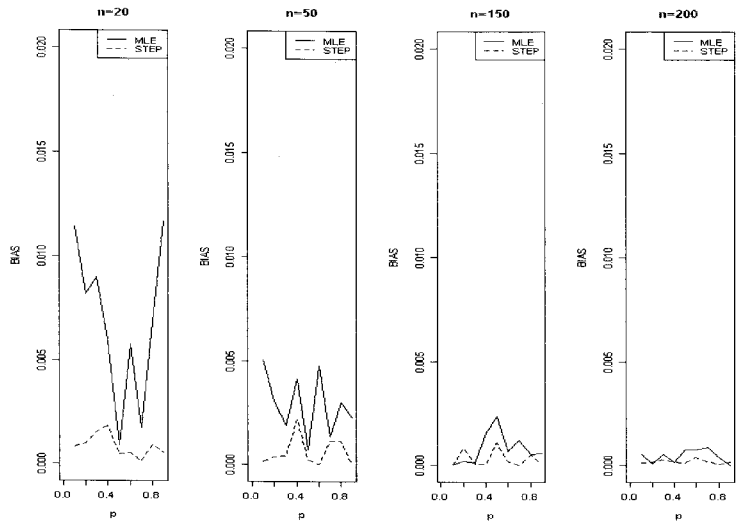

그림 1: *case 1에서의 MLE_{ext}과 MLE_{step}의 절대편향 비교*

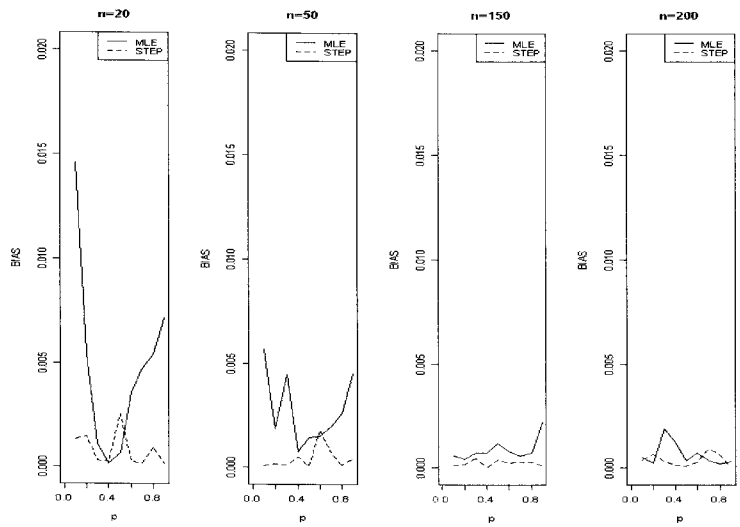

그림 2: *case 2에서의 MLE_{ext}과 MLE_{step}의 절대편향 비교*

서는 정확최대우도추정량을 그리고 이론적인 연구를 할 경우에는 단계별추출최대우도추정량을 사용하는 것이 타당하다고 판단된다.

참고 문헌

Blumenthal, S. (1968). Multinomial sampling with partially categorized data, *The Journal of the American Statistical Association*, **63**, 542–551.

Hocking, R. R. and Oxspring, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data, *The Journal of American Statistical Association*, **66**, 65–70.

Lee, S. E. and Park, C. J. (1999). A note on estimation of multinomial probabilities when some frequency counts are merged, *The Korean Communication in Statistics*, **6**, 327–336.

Park C. J., Shin, K. -I., Kim-Park, Y. and Lee, S. (2006). A note on Bayes estimates of multinomial probabilities under step-wise sampling, *The Journal of Applied Probability & Statistics*, **1**, 101–114.

2009년 9월 접수; 2009년 11월 채택

Comparison of Step-Wise and Exact Maximum Likelihood Estimations on Cell Probabilities of Contingency Table

Sang Eun Lee[a], Kee Hoon Kang[b], Seok-O Jeung[b], Key-Il Shin[1,b]

[a]Department of Applied Statistics, Kyonggi University
[b]Department of Statistics, Hankuk University of Foreign Studies

Abstract

In multinomial scheme with step-wise sampling, maximum likelihood estimates of multinomial probabilities are improved when some frequencies are merged. In this study, for cell probabilities in a I by J independent contingency tables, exact MLE and step-wise estimation methods are applied and the results are compared using MSE and Bias.

Keywords: Step-wise sampling, maximum likelihood estimation, contingency table, Newton-Raphson method.

This research was supported by the Hankuk University of Foreign Studies research fund 2009.

[1] Corresponding author: Professor, Department of Statistics, Hankuk University of Foreign Studies, Yonginsi, Kyonggy 449-791, Korea. E-mail: keyshin@hufs.ac.kr