응용통계연구 제 7권 2호
1994년 9월 pp. 269-277

최소카이제곱추정과 붓스트랩[1)]

정한영[2)], 이기원[3)], 구자용[4)]

요약

최소카이제곱추정에 의하여 구한 추정량의 표본분포를 붓스트랩으로 근사시켰을 때에도 정규근사와 최소한 동등함을 설명하고, 이 이론을 자궁경부암 조직에서 검출되는 란게르한스 세포의 출현률 추정에 이용하였다. 란게르한스 세포의 출현횟수를 포지티브 포아송 모형에 적합시켰으며, 추정된 출현률의 표준오차는 대표본 근사 및 붓스트랩을 이용하여 계산하였다. 두 방법 모두 비슷한 결과를 제공하였다.

1. 이론

θ를 모수로 하는 확률분포 $p(x;\theta)$로부터의 랜덤표본 $X_1, \cdots, X_n$ 에 대하여 j번째 쎌의 출현확률이 $\pi_j(\theta)$ $(\pi_1(\theta) + \cdots + \pi_k(\theta) = 1)$ 이 되도록 분류하여 각 쎌에 나타난 빈도를 $O_1, \cdots, O_k$ $(O_1 + \cdots + O_k = n)$라 하자. 이 때 θ의 최소 카이제곱추정량 $\hat{\theta}$은 피어슨 카이제곱상위(Pearson chi-square discrepancy)를 최소로 하며, 모형선택에서 많이 쓰이는 방식으로 표현하면 다음과 같다 (Linhart and Zucchini, 1986).

$$\hat{\theta} = \arg\min_{\theta} \sum_{j=1}^{k} \frac{\{O_j - E_j(\theta)\}^2}{E_j(\theta)}. \tag{1.1}$$

여기서, $E_j(\theta) = n \times \pi_j(\theta)$는 j 번째 쎌에서 기대되는 출현빈도이다. 혼동의 우려가 없으면, 합기호 내에서의 첨자는 생략하기로 한다. Rao(1957)는 이 문제에 있어서 θ 가 실수인 경우만을 고려하여도 충분함을 설명하고, 이렇게 구한 최소 카이제곱추정량의 대표본 정규근사 및 강일치성(strong consistency)등을 입증한 바 있다. 그 결과를 우리 문제에 응용하기 쉽도록 요약하면 다음과 같다.

적절한 조건 하에서 $n^{1/2}(\hat{\theta} - \theta)$ 의 표본분포는 표본의 크기가 늘어남에 따라 평균이 0이고 분산이 $\{\sum \pi'^{\,2}(\theta)/\pi(\theta)\}^{-1}$ 인 정규분포로 약수렴하고, 이 때 분산 추정량의 역수

1) 이 연구는 1993년도 학술진흥재단의 자유공모과제 학술연구조성비에 의하여 이루어졌음.
2) (200-702) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 교수.
3) (200-702) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 부교수.
4) (200-702) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과 조교수.

$\sum\pi'^{\,2}(\hat{\theta})/\pi(\hat{\theta})$는 $\sum\pi'^{\,2}(\theta)/\pi(\theta)$로 확률수렴한다. 여기서, π_j' 은 π_j 의 1차도함수이다.

따라서, (1.1) 의 해 $\hat{\theta}$ 의 표준오차는 다음의 식으로 근사된다.

$$SE_A(\hat{\theta}) \approx \sqrt{\{\sum\pi'^{\,2}(\hat{\theta})/\pi(\hat{\theta})\}^{-1}/n}. \tag{1.2}$$

한편, Efron(1979)에 의하여 도입되어 통계학의 많은 분야에 커다란 영향을 미친 바 있는 붓스트랩은 표준오차가 간단히 계산되지 않는 문제들에 대하여 컴퓨터 집중적인 해결방법을 제공해 주고 있다. Beran and Ducharme(1991), Hall(1992) 및 Efron and Tibshirani(1993)는 각각 그 이론적 측면에 대한 연구결과와 실제적 응용면을 살필 수 있는 좋은 참고 문헌이다. 이 붓스트랩을 본 문제에 적용시키면 다음과 같은 절차에 따라 최소 카이제곱추정량의 표본분포를 근사시킬 수 있게 된다.

(가) 붓스트랩 표본 $X_1^*, \cdots, X_n^*$를 적합된 분포 $p(x, \hat{\theta})$ 로부터 추출한다.

(나) (가)의 붓스트랩 랜덤표본으로부터 $O_1^*, \cdots, O_k^*$ 를 구한다.

(다) 붓스트랩 최소 카이제곱추정량 $\hat{\theta}^*$ 를 다음과 같이 구한다.

$$\hat{\theta}^* = \arg\min_{\theta} \sum_{j=1}^{k} \frac{\{O_j^* - E_j(\theta)\}^2}{E_j(\theta)}. \tag{1.3}$$

(라) (가)-(다)의 과정을 충분히 많은 횟수, 예를 들어, B회 반복하여 $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ 를 구한다.

(마) $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$ 의 표본분포함수(empirical cumulative distribution function)로 $\hat{\theta}$의 표본분포를 근사시킨다.

이 때, $\hat{\theta}$ 의 붓스트랩 표준오차는 다음의 식으로 근사된다 (Efron and Tibshirani, 1993).

$$SE_B(\hat{\theta}) \approx \sqrt{\sum_{l=1}^{B}(\hat{\theta}_l^* - \overline{\hat{\theta}}^*)^2/(B-1)}, \tag{1.4}$$

단, $\overline{\hat{\theta}}^* = \sum_{l=1}^{B} \hat{\theta}_l^*/B.$

(가)-(다)의 과정에서 계산되는 $\hat{\theta}^*$ 에 대하여 Rao(1957)의 정리와 Beran(1984)의 정리를 적용시키면 다음과 같이 붓스트랩 근사도 최소한 대표본 근사에 필적함을 보일 수 있다.

표본의 크기가 늘어남에 따라 $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ 의 분포는 평균이 0이고 분산이 $\{\sum\pi'^{\,2}(\theta)/\pi(\theta)\}^{-1}$인 정규분포로 약수렴하며, $\sum\pi'^{\,2}(\hat{\theta}^*)/\pi(\hat{\theta}^*)$는 $\sum\pi'^{\,2}(\theta)/\pi(\theta)$로 확률수렴한다.

2. 실제 자료에의 응용

2.1 자료의 개요

이기원(1992)은 자궁경부암을 대상으로 S-100단백을 이용하여 검출되는 란게르한스세포와 악성변화와의 연관성에 대하여 모형선택의 관점에서 분석한 바 있다. 피어슨 타입의 카이제곱 상위를 이용하였기 때문에 출현률은 최소 카이제곱법으로 추정하였으며, AIC 타입의 모형선택 기준에 의하여 최적 부모형을 선택하였다. 이 자료의 출처 및 내용에 대하여는 Lee, et. al.(1989) 및 이기원(1992)를 참조하면 된다.

<표 1> 본 연구에 사용된 자료

	+	++	+++	
Dysplasia (이형성증)	6	7	0	13
Carcinoma in situ(상피내암)	2	6	6	14
Invasive carcinoma (침윤)	1	3	6	10
계	9	16	12	37

본 연구에서는 이 자료 중 대조군을 제외한 자궁경부 이형성증, 상피내암 및 침윤성편평상피내암의 경우에 대하여 가능한 부모형들을 설정하여 포지티브 포아송 분포를 적합시키고 그 때 계산되는 출현률 추정량의 표준오차를 대표본 근사이론 및 붓스트랩에 근거하여 계산, 상호 비교하고자 한다 (표 1 참조). 대조군을 제외하는 이유는 실제 자료를 수집한 병원 스텝들의 의견 및 이기원(1992)에서의 모형선택 결과에 기인한다. 이기원(1992)에서의 기호를 따르는 한편 표현을 간결하게 하기 위하여 자궁경부 이형성증, 상피내암 및 침윤성편평상피내암 집단을 각각 2집단, 3집단, 및 4집단이라 하고, 각 집단에서의 란게르한스 세포의 출현률을 λ_2, λ_3, 및 λ_4 라 하자. 이 때 S-100단백질이 각 집단을 얼마나 잘 구별하는 지를 판단하는 데 쓰이는 부모형들로는 다음의 4가지를 들 수 있다.

A모형 ; $\lambda 2 < \lambda 3 < \lambda 4$.
C모형 ; $\lambda 2 = \lambda 3 < \lambda 4$.
D모형 ; $\lambda 2 < \lambda 3 = \lambda 4$.
E모형 ; $\lambda 2 = \lambda 3 = \lambda 4$.

각 모형이 의미하는 바는 자명하므로 설명을 생략한다. 먼저 이 자료를 적합시키는데 사용될 포지티브 포아송 분포, 즉, 무출현의 경우를 절단시킨 포아송 분포에 대하여 알아보자.

2.2. 포지티브 포아송 분포

David and Johnson(1952), Cohen(1954), Craig(1953), Johnson and Kotz(1969) 등은 포지티브 포아송 분포의 응용에 대하여 살펴볼 수 있는 참고 문헌들이며, Seber(1980)는 동물 모집단의 크기 추정에 응용되는 예들과 다른 많은 참고 문헌들을 제공하고 있다.

평균 출현률이 λ 인 포아송 분포의 경우 k 개의 출현확률 p_k 와 $k-1$ 개의 출현확률 p_{k-1} 사이에는 다음과 같은 관계가 성립한다.

$$p_k = \frac{\lambda}{k} p_{k-1}, \qquad k \geq 1. \tag{2.1}$$

여기서, 초기값 p_0는 $\exp(-\lambda)$로 주어진다. 이 관계식은 절단된 포아송 분포들에 대하여도 성립하는 데, 포지티브 포아송 분포의 경우에는 $k \geq 2$에 대하여 (2.1)이 성립하며, 이 때의 초기값 p_1은 $\lambda p_0/(1-p_0)$로 주어진다. 또한, 포지티브 포아송 분포를 따르는 확률변수 Y의 평균 및 분산은 다음과 같이 주어진다.

$$E(Y) = \frac{\lambda}{1-\exp(-\lambda)}, \quad Var(Y) = \frac{\lambda\{1-\exp(-\lambda)-\lambda\exp(-\lambda)\}}{\{1-\exp(-\lambda)\}^2}. \tag{2.2}$$

따라서, 포아송 분포와는 달리 분산이 평균보다 작음을 알 수 있으며, 평균은 포아송 출현률 λ의 단조증가함수임을 쉽게 알 수 있다.

(2.1)식을 이용하면 각 셀의 출현확률을 다음과 같이 구할 수 있다.

$$\begin{aligned} \pi(+) &= p_1(\lambda) \\ \pi(++) &= p_1(\lambda)(\lambda/2+\lambda^2/3!+\lambda^3/4!) \\ \pi(+++) &= 1-\pi(+)-\pi(++). \end{aligned} \tag{2.3}$$

한편, 추정출현률의 표준오차를 계산하기 위해서는 (2.3)의 1차도함수들을 계산할 필요가 생기는 데, (2.1)로부터 p_k 의 1차도함수 p_k'에 대하여 다음과 같은 관계식을 얻을 수 있다.

$$p_k' = p_{k-1} - p_k/(1-p_0), \qquad k \geq 2. \tag{2.4}$$

여기서, 초기값 p_1'은 $(1-\lambda-p_0)p_0/(1-p_0)^2$로 주어진다. 이 관계식을 이용하여 각 셀 출현확률의 1차도함수가 쉽게 구해진다.

뉴튼-랍슨 방법에 의하여 반복적으로 출현률을 계산할 때에는 2차도함수까지도 요구된다. 여

기서 p_k'' 를 p_k의 출현률에 관한 2차도함수라 하고, 표현을 보다 간결히 하기 위하여 $\rho = p_0/(1-p_0)$라 하면 다음 식이 성립한다.

$$p_k'' = p_{k-1}' - (p_k' - \rho p_k)/(1-p_0), \qquad k \geq 2. \tag{2.5}$$

여기서, 초기값 p_1''은 $-\rho/(1-p_0) - (p_1' - \rho p_1)/(1-p_0)$로 주어진다.

2.3. 출현률의 추정방법

본 연구에서는 (1.1)의 해를 구하기 위하여 뉴튼-랍슨에 근거한 반복적방법(iterative method)과 이기원(1992)에서 소개한 바 있는 탐색방법(search method)을 모두 사용하여 비교하였다. 그 결과 일정한 탐색구간을 설정하여 그 구간 내에서 해를 찾는 것이 보다 빠른 결과를 제공해 주기 때문에 붓스트랩 작업 시에는 탐색방법을 채택하게 되었다.

탐색방법의 경우, 탐색구간을 정할 때 경험적으로 충분히 큰 구간을 미리 설정하여 그 구간 내에서 해를 찾는 고정격자 탐색방법(fixed grid search method)과 자료에 의존하여 탐색구간의 상, 하한을 정하는 랜덤격자 탐색방법(random grid search method)등을 생각해 볼 수 있다. 두 방법을 비교해본 결과 랜덤격자 탐색방법으로 상, 하한을 정할 경우 탐색구간의 폭은 짧게 잡을 수 있으나, 실제 총소요시간이 고정격자 탐색방법보다 오래 걸리고, 붓스트랩 작업 시 탐색구간을 반복 조정해 주어야 할 경우가 종종 발생하여 결국 경험적으로 얻은 충분히 큰 탐색구간을 이용하게 되었다. 다음은 각 방법의 개요이다.

(가) 고정격자 탐색방법

가장 간단한 방법이면서도 계산속도가 빨라서 붓스트랩에 사용한 방법이다. 경험적으로 정한 [0.01, 10] 을 그 탐색구간으로 하되 격자 간격을 0.01씩으로 하였다.

(나) 랜덤격자 탐색방법

고정격자 탐색방법이 경험적으로 충분히 큰 탐색구간을 설정하는 데 비하여, 자료에 따라 탐색구간의 상, 하한을 결정하는 방법으로 탐색구간의 폭을 줄이는 데 목적이 있다. 먼저, 탐색구간의 중점을 자료로부터 계산해야 하는데, 가장 간단한 방법은 "+"에서 실제 관찰된 횟수와 그 셀에서의 기대횟수를 등식으로 하여 출현률의 초기값을 구하는 것이다. 이 방정식을 뉴튼-랍슨 방법으로 풀면, r번째 반복 과정에서의 식은 다음과 같이 주어진다.

$$\lambda_{(r)} = \lambda_{(r-1)} - \frac{\lambda_{(r-1)}\exp(-\lambda_{(r-1)})/\{1-\exp(-\lambda_{(r-1)}) - O(+)/n\}}{\{1-\lambda_{(r-1)} - \exp(-\lambda_{(r-1)})\}/\{1-\exp(-\lambda_{(r-1)})\}^2}. \tag{2.6}$$

이와 같은 방법으로 구한 출현률의 초기추정량을 $\hat{\lambda}_0$ 라 하자. 이 문제의 경우에는 단순히 탐색구간의 상, 하한을 정하면 되므로, 포아송 분포의 경우에 평균과 분산이 같은 점을 이용하여, 탐색구간의 상한은 $\hat{\lambda}_0 + k \times \sqrt{\hat{\lambda}_0/n}$, 하한은 $\max\{0.01,\ \hat{\lambda}_0 - k \times \sqrt{\hat{\lambda}_0/n}\}$로 정하였다. 배수 k를 4로 하여 붓스트랩을 수행시켜 본 결과 구간을 더 늘려야 할 경우가 가끔 발생하여 그 값

을 6으로 하였더니 더 이상 그런 경우가 발생하지 않았다.

(다)뉴튼-랍슨 방법

(1.1)의 해를 구하기 위하여서는 그의 1차도함수 및 2차도함수를 필요로 한다. (2.4) 및 (2.5)를 이용하여 r 번째 단계에서의 반복식을 구하면 다음과 같다.

$$\lambda_{(r)} = \lambda_{(r-1)} - \frac{\sum O^2 E' / E^2}{\sum O^2 (E''/E^2 - 2E'^2/E^3)} \tag{2.7}$$

여기서, E' 과 E'' 은 각각 쎌 기대빈도의 출현률에 관한 1차도함수 및 2차도함수를 나타낸다. 이 때의 수렴 기준은 탐색방법에서의 격자 간격이 0.01인 만큼 이와 대등한 조건을 주기 위하여 전 단계에서의 출현률 값과의 차이가 0.001 이내이면 수렴하는 것으로 간주하였다.

표 2는 자궁경부 이형성증의 경우에 대하여 각 방법을 이용하여 계산하는 데 걸리는 평균시간을 비교한 것이다. 20회 씩 수행하였으며 양끝 5% 씩을 잘라낸 나머지 관찰값들의 평균시간을 취하였다. 절삭하지않은 전체 관찰값들로부터 얻은 표준오차는 대략 0.1초정도이었다. 수행시간은 S-Plus의 dos.time함수로 측정하였으며 486 SLC (25/50 MHz) 칩을 장착한 IBM ThinkPad 720 노트북 컴퓨터가 사용되었다.

<표 2> 출현률 추정방법들의 속도 비교

평균 소요시간	고정격자 탐색방법	랜덤격자 탐색방법	뉴톤-랍슨 방법
λ_2	1.58	1.20	1.92
λ_3	1.73	2.24	4.95
λ_4	1.76	1.52	4.03
$\lambda_2 = \lambda_3$	1.82	1.82	5.82
$\lambda_3 = \lambda_4$	1.96	2.14	5.45
$\lambda_2 = \lambda_3 = \lambda_4$	1.81	2.63	5.77

뉴튼-랍슨 방법이 탐색방법들에 비하여 계산시간이 현저히 오래 걸림을 알 수 있다. 또한 랜덤격자 탐색방법의 경우, 앞에서 지적한 바와 같이 탐색구간이 해를 놓치는 경우가 가끔 발생하고 실제 붓스트랩 작업 시에는 고정격자 탐색방법보다 훨씬 많은 시간이 소요되어 고정격자 탐색방법을 주로 사용하게 되었다. 예를 들어, λ_4의 경우 붓스트랩 반복수를 200으로 하였을 때, 고정격자 탐색방법에 근거한 붓스트랩은 384초, 랜덤격자 탐색방법에 근거한 붓스트랩은 543초 걸리는 것을 관찰할 수 있다. 출현률의 추정값, 대표본 근사이론에 근거한 표준오차 및 최소 카이제곱값은 이미 이기원(1992)에 실려 있으며, 본 연구에서는 붓스트랩과의 비교를 위하여 출현률의 추정값과 각 방법에 의한 표준오차들만을 수록하였다.

2.4 출현률의 추정값과 표준오차

2.3절에서 소개한 방법으로 출현률의 추정값을 구한 후에는, (1.2)식을 이용하여 대표본 근사이론에 근거한 표준오차를 계산할 수 있다. (1.3)식 및 (1.4)식에 근거한 붓스트랩의 경우에는 반복수를 50, 100 및 200으로 하여 그 변화를 살필 수 있도록 하였다. 이러한 붓스트랩 반복수의 선택은 Efron and Tibshirani(1993)에 근거한다.

표 3은 이 결과를 기록한 것이다. 이 표에서 SE_A는 대표본 근사이론에 의해 계산한 표준오차이고, $SE_{B,50}$, $SE_{B,100}$ 및 $SE_{B,200}$은 각각 반복횟수를 50, 100 및 200으로 하여 계산한 붓스트랩 표준오차들이다. 양 방법에 의하여 구한 표준오차들이 상당히 근사하여 붓스트랩 근사가 델타방법에 필적함을 알 수 있다.

<표 3> 출현률의 추정값과 표준오차

모 형	추정값	SE_A	$SE_{B,50}$	$SE_{B,100}$	$SE_{B,200}$
A	λ_2 = 1.35 λ_3 = 3.91 λ_4 = 4.70	0.4376 0.6313 0.8421	0.4811 0.5990 0.9184	0.4307 0.6179 0.8630	0.4726 0.6255 0.9348
C	$\lambda_2=\lambda_3$ = 2.72	0.3883	0.4098	0.3777	0.3291
D	$\lambda_3=\lambda_4$ = 4.23	0.5052	0.5019	0.5375	0.5137
E	$\lambda_2=\lambda_3=\lambda_4$ = 3.20	0.3529	0.3004	0.3723	0.3296

3. 맺음말

어떤 추정량의 표준오차를 계산하는 데 있어서, 대표본 근사이론에 바탕을 두고 있는 델타방법은 그 포괄적인 특성에 기인하여 많이 활용되고 있으나 일정 수준 이상의 해석적 계산능력이 항상 요구된다. 반면 방대한 양의 단순 계산을 쉽고 저렴하게 처리할 수 있는 컴퓨터의 대중화와 S-Plus 류의 전문적 계산 소프트웨어의 발달은 붓스트랩과 같이 컴퓨터 집약적인 통계기법의 발전에 크게 공헌하였다. 따라서,해석적 표현을 얻기 힘든 통계문제의 해결에도 많은 기여를 할 것으로 보인다. 본 연구에서는 붓스트랩을 이용한 표준오차 계산이 대표본 근사이론에 근거한 방법과 대등한 결과를 제공해 줄 수 있음을 보였다.

모든 계산은 S-plus함수를 개발하여 수행하였으며, 저자들로부터 입수가능함을 밝혀둔다.

참고문헌

[1] 이기원 (1992). 란게르한스 세포의 출현횟수에 대한 통계적 고찰, 「응용통계연구」 제5권 2호, 271-282.

[2] Beran, R.J. (1984). Bootstrap methods in statistics, *Jber. d. dt. Math. Verein.* Vol. 86, 14-30.

[3] Beran, R.J. and Ducharme, G. R. (1991). *Asymptotic theory for the Bootstrap methods in statistics*, Centre de Reserches Mathematiques, University of Montreal.

[4] Cohen, A.C. (1954). Estimation of the Poisson parameter from truncated samples and from censored samples, *Journal of the American Statistical Association,* Vol. 49, 158-168.

[5] Craig, C.C. (1953). On the utilization of marked specimens in estimating populations of flying insects, *Biometrika*, Vol. 40, 170-176.

[6] David, F.N. and Johnson, N.L. (1952). The truncated Poisson, *Biometrics*, Vol. 8, 275-285.

[7] Efron, B. (1979). Bootstrap methods, *Annals of Statistics*, Vol. 7, 1-26.

[8] Efron, B. and Tibshirani, R. (1993). *An introduction to the Bootstrap*, Chapman & Hall, New York.

[9] Hall, P.G. (1992). *Bootstrap and Edgeworth expansion*, Springer-Verlag, New York.

[10] Johnson, N.L. and Kotz, S. (1969). *Discrete distributions,* Wiley, New York.

[11] Lee, M.C., Park, K.M., Kang, G., Lee, D.Y., Shin, H.S., and Park, Y.E.(1989). An immunohistochemical study on the subpopulation of Langerhans cells in cervical carcinoma using S-100 protein, *Hallym University Journal Natural Sciences and Medicine,* Vol. 7, 159-168.

[12] Linhart, H., and Zucchini, W. (1986). *Model selection*, Wiley : New York.

[13] Rao, C.R.(1957). Theory of the method of estimation by minimum chi-square, *Bulletin of the International Statistical Instistitute*, Vol. 35(2), 25-32.

[14] Seber, G.A.F.(1980), *The Estimation of animal abundance,* Charles Griffin & Co., London.

Minimum Chi-square Estimation and the Bootstrap [1)]

Han-Yeoung Chung[2)], Kee-Won Lee[3)], Ja-Yong Koo[4)]

Abstract

Bootstrap approximation is compared with ordinary asymptotic method in the context of minimum chi-square estimation through application in a real problem. Fixed interval search method is shown to be superior over a random interval search method or Newton-Raphson method. All the procedures are implemented by S-Plus functions.

---

1) This research was supported in part by Non-Directed Research Fund, Korea Research Foundation, 1993.

2) Professor, Department of Statistics, Hallym University, Chunchon, 200-702 KOREA.

3) Associate Professor, Department of Statistics, Hallym University, Chunchon, 200-702 KOREA.

4) Assistant Professor, Department of Statistics, Hallym University, Chunchon, 200-702 KOREA.