Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 1, pp. 29 ~ 36

안부점근사를 이용한 승산비에 대한 점근적 추론

나종화[1]

요약

분할표 분석에서 승산비(odds ratio)에 대한 추론은 중요하다. 이에 대한 정확한 추론은 비중심초기하(noncentral hypergeometric) 분포의 누적확률등의 계산이 요구되어 표본의 크기가 클 경우 많은 양의 계산과 계산시간이 요구되므로 StatXact 등의 프로그램을 이용하는 것이 일반적이다. 본 논문에서는 정확한 추론에 대한 대안적 방법으로 안부점 근사(saddlepoint approximation)의 결과를 이용한 점근적 추론법을 제시하였다. 이 방법은 비교적 소표본의 경우에도 정확한 추론의 결과와 일치하며, 기존의 정규근사를 이용한 방법에 비해 매우 뛰어난 정확도를 유지함을 예제를 통해 확인하였다.

주제어: 승산비, 안부점근사, 점근적 추론

1. 서론

다음과 같은 p-차원 모수를 가지는 지수족 모형을 생각하자.

$$f_{S,T}(s,t;\psi,\nu) = exp\{\psi s + \nu' t - K(\psi,\nu) + h(s,t)\}. \tag{1}$$

여기서 ψ는 관심 있는 모수이고 $\nu = (\nu_1, \cdots, \nu_{p-1})$은 장애(nuisance) 모수이다. Fisher(1934)에 의하면 모수 ψ에 관한 정확한 추론은 다음의 조건부 분포에 기초하는 것이 타당하다.

$$f_{S|T=t}(s;\psi) = exp\{\psi s - K^*(\psi;t) + h(s,t)\}. \tag{2}$$

예를 들어, 모수 ψ에 대한 $100(1-\alpha)\%$ 신뢰 상한 ψ_U는 $Pr\{S \geq s|T=t;\psi_U\} = \alpha$를 만족하는 값으로 주어진다. 일반적으로 조건부 밀도함수 (2)가 구체적인 형태로 표현되는 경우는 드물기 때문에 모수 ψ에 대한 정확한 조건부적 추론은 특별한 문제에서만 제한적으로 사용될 수 있다. 본 논문에서는 모수 ψ의 추론에 사용되는 다음의 조건부 누적확률

$$Pr\{S \leq s|T=t;\psi\} \tag{3}$$

[1]충북 청주시 흥덕구 개신동 충북대학교 통계학과 조교수

에 대한 안부점 근사식을 소개하고, 이에 기초한 통계적 추론 문제로서 다항적 모형을 가정한 2×2 분할표에서 승산비에 대한 점근적 추론법을 제시하였다.

한편, 2×2 분할표에서 각 칸의 기대돗수와 관측돗수를 각각 m_{ij} $(i,j = 1,2)$ 와 x_{ij} $(i,j = 1,2)$로 표현하고 각 칸에 포함될 확률을 P_{ij} $(i = 1,2)$라 할 때, 대수 승산비(log odds ratio) $\psi = log(P_{12}P_{21}/P_{11}P_{22})$ 에 대한 기존의 $100(1-\alpha)\%$ 근사 신뢰구간은 다음의 ψ에 대한 최우추정량 $\hat{\psi} = log(X_{12}X_{21}/X_{11}X_{22})$ 의 근사분포 즉,

$$\hat{\psi} \sim ASN\left(\psi,\ \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{m_{ij}}\right) \tag{4}$$

으로부터 다음과 같이 주어진다. (여기서 ASN(·)은 점근정규분포를 나타낸다.)

$$\hat{\psi} \pm z_{\alpha/2}\sqrt{\hat{V}ar(\hat{\psi})}. \tag{5}$$

식(5)에 기초한 점근적 추론의 결과는 표본의 크기가 충분히 크지 않은 경우에는 정확한 추론의 결과와 매우 다른 값으로 주어지며, 정확한 추론 역시 비중심초기하 분포의 누적확률 등의 계산이 요구되어 표본이 커짐에 따라 계산상에 어려움이 발생한다.

본 연구에서는 안부점 근사에 기초한 최근의 이론을 바탕으로 승산비에 대한 새로운 근사적 추론법을 제시하였고 예제를 통해 이 방법의 우수성을 확인하였다. 2절에서는 식(3)의 조건부 누적확률에 대하여 안부점 근사법을 간략히 소개하고 3절에서는 2×2분할표에서 승산비에 대한 점근적 추론 문제를 다루었으며 4절에서는 예제를 통하여 제시한 방법의 우수함을 확인하였다.

2. 조건부 분포함수에 대한 안부점 근사

먼저 모수 ψ에 대한 점근적 추론에 다음의 우도비 또는 이탈도(deviance)통계량

$$W(\psi) = 2\{\ell(\hat{\psi},\hat{\nu}) - \ell(\psi,\hat{\nu}_\psi)\} \tag{6}$$

이 점근적으로 자유도가 1인 카이자승분포를 따른다는 사실을 이용할 수 있다. 여기서 $(\hat{\psi},\hat{\nu})$는 비제약적(unrestricted) 최우추정량이고 $\hat{\nu}_\psi$ 은 ψ가 주어진 상황(가설에서)하의 모수 ν에 대한 최우추정량을 의미한다. 식(6)은 프로화일(profile) 대수-우도함수(log-likelihood function)라 불리우는 다음의 식

$$\ell_p(\psi) = \ell(\psi,\ \ \hat{\nu}_\psi) \tag{7}$$

을 이용하면 다음의 형태로 재표현 될 수 있다.

$$W(\psi) = 2\{\ell_p(\hat{\psi}) - \ell_p(\psi)\}. \tag{8}$$

만약 ν의 차원이 고정되면, 식(7)은 1차원 모수에 대한 우도함수의 역할을 한다. 따라서 방향성 이탈도(directed deviance) 또는 부호화 우도비(signed LR) 통계량으로 불리우는 다음의 통계량

$$Z_{dev}(s,t;\psi) = sgn(\hat{\psi}-\psi)\{W(\psi)\}^{\frac{1}{2}} \tag{9}$$

이 점근적으로 정규분포를 따른다는 사실로부터 모수 ψ에 대한 신뢰구간을 구할 수 있게 된다. 그러나 장애 모수가 많거나 표본의 크기가 작을 때는 프로화일 우도함수가 조건부 우도함수로 부터 크게 차이가 나는 경향이 있어 이를 수정한 수정된 프로화일(adjusted profile) 우도함수를 Barndorff-Nielsen과 Cox(1979)가 다음과 같이 제안하였다.

$$\ell_a(\psi) = \ell_p(\psi) + log|I_\nu(\psi,\hat{\nu}_\psi)|^{\frac{1}{2}}. \tag{10}$$

여기서 $I_\nu(\psi,\hat{\nu}_\psi)$는 ψ가 주어질 때 ν에 대한 Fisher-정보량(Fisher information)으로 ψ와 $\hat{\nu}_\psi$에서 계산된 값이다. Pierce 와 Peters(1992)는 식(9)에 대응하는 이탈도 통계량 Z_{adev}를 다음의 식으로 근사할 수 있음을 보였다.

$$Z_{adev} \simeq Z_{dev} + log\rho/Z_{dev}. \tag{11}$$

여기서

$$\rho = |I_\nu(\hat{\psi},\hat{\nu})|^{\frac{1}{2}}/|I_\nu(\psi,\hat{\nu}_\psi)|^{\frac{1}{2}} \tag{12}$$

이다.

한편, 식(3)에 대한 근사로는 Barndorff-Nielsen(1986), Skovgaard(1987), Fraser(1990,1991) 등이 있으며 이들의 근사식은 모두 연속형의 모형에만 적용될 수 있는 형태이다. Pierce와 Peters(1992)는 이산형의 자료에도 적용될 수 있는 식(3)에 대한 연속형 수정(continuity correction) 을 고려한 여러 가지 근사식을 제안하였다. 이 근사식들은 모두 안부점근사에 기초한 식들로서 Daniels(1987)의 기법을 사용하여 유도된 결과로써, 본 연구에서는 다음의 근사식을 이용하여 승산비에 대한 추론을 실시하고자 한다.

$$Pr(S \le s|T=t;\psi) \simeq \Phi\{Z^{*+}_{dev} + log\{g(\hat{\psi}^{+}-\psi)\}/Z^{+}_{dev}\}. \tag{13}$$

여기서 Z^{*+}_{dev}, $\hat{\psi}^{+}$와 Z^{+}_{dev}는 s 대신 $s+0.5$로 연속성 수정된 자료로부터 계산되는 양이며 Z^{*}_{dev}의 정의는 다음과 같다.

$$Z^{*}_{dev} = Z_{dev} - log(Z_{dev}/\rho Z_{wald})/Z_{dev}. \tag{14}$$

또한 Z_{dev}는 식(9)와 같이 주어지고 Z_{wald}는

$$Z_{wald} = (\hat{\psi}-\psi)|I_\theta(\hat{\theta})|^{\frac{1}{2}}/|I_\nu(\hat{\psi},\hat{\nu})|^{\frac{1}{2}} \tag{15}$$

으로 주어지며 $\hat{\theta}=(\hat{\psi},\hat{\nu})$이고 $I_\nu(\hat{\psi},\hat{\nu})$은 $I_\theta(\hat{\theta})$의 $(p-1)\times(p-1)$차원 부행렬(submatrix)을 의미한다. 식(13)에서 함수 $g(\cdot)$에 대한 정의는 다음과 같다.

$$g(\hat{\psi}-\psi) = [exp\{(\hat{\psi}-\psi)/2\} - exp\{-(\hat{\psi}-\psi)/2\}]/(\hat{\psi}-\psi). \tag{16}$$

3. 승산비에 대한 점근적 추론

다음의 2×2분할표를 생각하자.

X_{11}	X_{12}	$X_{1\cdot}$
X_{21}	X_{22}	$X_{2\cdot}$
$X_{\cdot 1}$	$X_{\cdot 2}$	n

여기서 각 칸에 포함될 잠재적 확률을 $P_{ij}, i,j = 1,2$라 하고 모수 ψ와 ν을 다음과 같이 정의하자.

$$\psi = log(P_{12}P_{21}/P_{11}P_{22}). \tag{17}$$

$$(\nu_1, \nu_2) = \{log(P_{22}/P_{12}), log(P_{22}/P_{21})\}. \tag{18}$$

다항분포로 부터의 표본추출을 가정할 때 대수-우도함수는 다음과 같이 주어짐을 보일 수 있다.

$$\begin{aligned} \ell(\theta) &= \ell(\psi, \nu) \\ &\propto -\psi x_{11} - \nu_1 x_{1\cdot} - \nu_2 x_{\cdot 1} + n\{\nu_2 - log(1 + e^{\nu_2} + e^{\nu_2 - \nu_1} + e^{-\nu_1 - \psi})\}. \end{aligned} \tag{19}$$

여기서 $\theta = (\psi, \nu)$이다. 식(19)로 부터 θ에 대한 최우추정치는 다음과 같다.

$$(\hat{\psi}, \hat{\nu}) = \{log(x_{12}x_{21}/x_{11}x_{22}),\ \ log(x_{22}/x_{12}),\ \ log(x_{22}/x_{21})\}. \tag{20}$$

고정된 ψ값에 대한 ν의 제한된 최우추정량 $\hat{\nu}_\psi$도 수치적(또는 이론적) 방법으로 쉽게 구해질 수 있다. 모수 ψ에 대한 추론에 사용되는 프로화일 대수-우도함수 $\ell_p(\psi)$와 방향성 이탈도 함수 Z_{dev}는 각각 (7)와 (9)식으로 주어진다. 또한 식(12)와 (15)의 ρ와 Z_{wald}값은 다음의 대수-우도함수에 대한 다음의 2차 미분값으로 부터 쉽게 구해진다.

$$\begin{aligned} \ell_{\psi\psi} &= -x_{11} + nB/A, \quad \ell_{\nu_1\nu_1} = -nB(A + A')/A^2, \\ \ell_{\nu_2\nu_2} &= -nC(1 - C/A)/A, \quad \ell_{\psi\nu_1} = \ell_{\nu_1\psi} = -nD(A - B)/A^2, \\ \ell_{\psi\nu_2} &= \ell_{\nu_2\psi} = -nCD/A^2, \quad \ell_{\nu_1\nu_2} = \ell_{\nu_2\nu_1} = -n(BC - AE)/A^2. \end{aligned} \tag{21}$$

여기서 l_{ab}의 표현은 대수-우도함수의 a와 b에 대한 2차 미분값을 의미한다. 위 식에서 사용된 A, A', B, C, D 와 E는 다음의 식으로 주어진다.

$$\begin{aligned} A &= 1 + e^{\nu_2} + D + E, \quad A' = (\partial/\partial\nu_1)A = 1 - D - E, \\ B &= D + E, \quad C = e^{\nu_2} + E, \quad D = e^{-\nu_1 - \psi}, \quad E = e^{\nu_2 - \nu_1}. \end{aligned} \tag{22}$$

위의 결과들을 식(13)에 적용하면 ψ의 추론에 사용되는 다음의 확률

$$Pr\{X_{11} \le x_{11} | (X_{1\cdot},\ \ X_{\cdot 1}) = (x_{1\cdot},\ \ x_{\cdot 1}); \psi\} \tag{23}$$

에 대한 근사를 실시할 수 있다. 식(13)의 계산에 사용된 모든 통계량들은 연속성 수정이 고려된 $x_{11}+0.5$의 값에 대한 결과이다. 다양한 ψ값에 대한 식(13)의 근사값으로 부터 모수 ψ에 대한 점근적 (또는 근사적) 신뢰구간을 형성할 수 있으며 본 논문에서 제시한 근사적 방법의 정확성을 알아보기 위해 4절에서는 정확한 계산이 가능한 실제자료의 예제를 통하여 비교해 보았다. 식(23)의 정확한 값은 다음의 비중심초기하 분포의 누적확률로 부터 구해진다.

$$\begin{aligned} Pr\{X_{11} \le x_{11} | (X_{1\cdot},\ X_{\cdot 1}) &= (x_{1\cdot},\ x_{\cdot 1}); \psi\} \\ &= \frac{\sum_{s=0}^{x_{11}} \binom{x_{1\cdot}}{s}\binom{x_{2\cdot}}{x_{\cdot 1}-s}(e^{\psi})^{x_{\cdot 1}-s}}{\sum_{t=0}^{x_{\cdot 1}} \binom{x_{1\cdot}}{t}\binom{x_{2\cdot}}{x_{\cdot 1}-t}(e^{\psi})^{x_{\cdot 1}-t}}. \end{aligned} \tag{24}$$

여기서 x_{11}은 X_{11}의 관찰빈도이다.

4. 예제를 통한 비교

Fisher(1934, 1962)에서 논의된 다음의 자료를 생각하자.

표 1: 동일성을 가지는 쌍둥이의 범죄의 유형

유형	판결		계
	유죄	무죄	
이란성	2	15	17
일반성	10	3	13
계	12	18	30

위 자료에 대해서 본 논문에서 제시한 안부점 근사를 실시한 결과는 표 2와 같다. 이 결과로 부터 안부점 근사의 결과가 비교적 소표본의 경우에도 정확도가 뛰어나며, 특히 통계적 추론에 사용되는 꼬리부분(tail part)의 영역에서도 정확도가 유지됨을 알 수 있다.

또한, 모수 ψ에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음의 관계식을 만족하는 ψ_U 와 ψ_L 로 부터 쉽게 구해진다.

$$P(\psi_U) = 1 - \alpha/2, \qquad P(\psi_L) = \alpha/2. \tag{25}$$

여기에서 $P(\cdot)$는 식(23)에 대한 근사(approximate) 또는 정확한(exact) 분포함수를 의미한다. 표 1의 자료로 부터 ψ에 대한 90% 신뢰구간을 비교하면 식(13)을 이용한 근사신뢰구간은(1.42 , 4.07)로 주어지고 식(24)를 이용한 정확한 신뢰구간은 (1.26 , 4.11)로 주어진다. 본 예제에서 신뢰 상한에서는 상당히 정확한 근사가 이루어지는 반면 신뢰하한에서는 약간의 차이를 보이고 있으나 심각한 차이는 아닌 것으로 판단된다. 한편, 식(5)로 부터 ψ에 대한 90% 근사신뢰구간은 (1.57 , 4.86)으로 주어지며, 이는 정확한 신뢰구간과는 상당한 차이를 보이고 있다. 따라서 본 연구에서 제시한 새로운 방법을 통하여 승산비에 대한 더욱 정확

표 2: 승산비에 대한 분포함수의 근사

ψ	Exact	Approximate	ψ	Exact	Approximate
1.0	0.0219338	0.0240353	4.6	0.9829078	0.9829072
1.2	0.0259155	0.0430338	4.8	0.9893613	0.9890497
1.4	0.0475503	0.0722718	5.0	0.9934944	0.9931168
1.6	0.0787138	0.1141284	5.2	0.9960842	0.9957466
1.8	0.1196426	0.1700286	5.4	0.9976755	0.9974118
2.0	0.1680331	0.2398238	5.6	0.9986367	0.9984466
2.2	0.2167528	0.3215011	5.8	0.9992090	0.9990790
2.4	0.2454820	0.4113288	6.0	0.9995451	0.9994599
3.6	0.8690261	0.8912442	6.2	0.9997405	0.9996863
3.8	0.9082975	0.9200117	6.4	0.9998529	0.9998193
4.0	0.9375162	0.9432907	6.6	0.9999170	0.9998967
4.2	0.9584958	0.9609979	6.8	0.9999534	0.9999413
4.4	0.9730753	0.9738755	7.0	0.9999740	0.9999669

한 점근적 추론이 가능하며, 기존의 정규근사에 기초한 점근적 추론의 대안으로 사용할 수 있음을 알 수 있다. 또한, 제시된 방법은 표본의 크기가 매우 큰 경우에 대해서도 기존의 정확한 추론에서 요구되는 많은 양의 계산과정을 피할 수 있어 실제의 응용문제에 효과적으로 사용될 수 있다.

5. 결론

조건부 분포함수에 대한 근사식을 바탕으로 분할표 분석에서 승산비에 대한 점근적 추론법을 제시하였다. 안부점 근사법에 기초한 이 방법은 정확한(exact) 추론시에 요구되는 비중심초기하 분포의 누적확률에 대한 계산을 피할 수 있고, 기존의 근사적 방법 보다는 더욱 정밀한 결과를 제공함을 예제를 통해 확인하였다. 특히, 본 논문에서 제시한 방법은 표본의 크기가 작거나 적당히 큰 경우에도 상당히 정확한 근사값을 제공한다.

참고문헌

1. Barndorff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika*, 73(2), 307-322.

2. Barndorff-Nielsen, O. E. and Cox, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *Journal of the Royal Statistical Society*, Series B, 41(3), 279-312.

3. Daniels, H. E. (1987). Tail probability approximations. *International Statistical Review*, 55(1), 37-48.

4. Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Statistical Society of London*, Series A, 144, 285-307.

5. Fisher, R. A. (1962). Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, 4, 41.

6. Fraser, D. A. S. (1990). Tail probabilities from observed likelihood. *Biometrika*, 77, 65-76.

7. Fraser, D. A. S. (1991). Statistical inference : Likelihood to significance. *Journal of the American Statistical Association*, 86(414), 258-265.

8. Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society*, Series B, 54(3), 701-737.

9. Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability*, 24, 875-887.

Asymptotic Inference on the Odds Ratio via Saddlepoint Method

Jonghwa Na [2]

Abstract

We propose a new method of asymptotic inference on the odds ratio (or cross-product ratio) in 2×2 contingency table. Saddlepoint approximations to the conditional tail probability are used in this procedure. We assess the accuracy of the suggested method by comparing with the exact one. To obtain the exact values, we need very complicated calculations containing the cumulative probabilities of non-central hypergeometric distribution. The suggested method in this paper is very accurate even for small or moderate sample sizes as well as simple and easy to use. Example with a real data is also considered.

Key Words and Phrases: Odds Ratio, Saddlepoint Approximation, Asymptotic Inference.

[2] Assistant Professor, Dept. of Statistics, Chungbuk National University of Cheong-Ju, Chungbuk, 361-763.