

## Logit Confidence Intervals Using Pseudo-Bayes Estimators for the Common Odds Ratio in $2 \times 2 \times K$ Contingency Tables

Donguk Kim<sup>1)</sup>, Eunhee Chun<sup>2)</sup>

### Abstract

We investigate logit confidence intervals for the odds ratio based on the delta method. These intervals are constructed using pseudo-Bayes estimators. The Gart method and Agresti method smooth the observed counts toward the model of equiprobability and independence, respectively. We obtain better coverage probability by smoothing the observed counts toward the pseudo-Bayes estimators in  $2 \times 2$  table. We also improve logit confidence intervals in  $2 \times 2 \times K$  tables by generalizing these ideas. Utilizing pseudo-Bayes estimators, we obtain better coverage probability by smoothing the observed counts toward the conditional independence model, no three-factor interaction model and saturated model in  $2 \times 2 \times K$  tables.

*Keywords:* Confidence interval; Coverage probability; Delta method; Odds ratio; Pseudo-Bayes estimator; Small sample;  $2 \times 2 \times K$  table.

### 1. 서론

분할표 분석에서 오즈비(odds ratio)는 매우 중요한 모수이며 이에 대한 신뢰구간 구축은 많이 연구되어 왔다.  $2 \times 2$  분할표에서 칸 확률이  $\{p_{ij}, i=1, 2, j=1, 2\}$  일 때 오즈비는  $\theta = p_{11} p_{22} / p_{12} p_{21}$  이며 오즈비의 추정량은  $\hat{\theta} = \hat{p}_{11} \hat{p}_{22} / \hat{p}_{12} \hat{p}_{21}$  이다. 포아송 분포, 다항 분포, 독립적 이항 분포를 따르는  $\{n_{ij}\}$ 에 대하여  $\theta$ 의 추정량에 양수  $c$ 를 사용하여  $\hat{\mu}_{ij} = n_{ij} + c$ 를 사용할 수 있다. 분할표에서  $n_{ij} = 0$ 인 경우 표본오즈비  $\hat{\theta}$ 는 0이거나 정의되지 않으므로 양수  $c$ 를 사용할 경우 이 문제를 해결할 수 있다. 표본크기가 작은 경우 오즈비는 비대칭 형태로 한쪽으로 치우쳐진 분포를 하나 수정된 추정량  $\hat{\theta}$ 에 로그변환을 하면 정규근사가 더 잘 된다.  $\log(\hat{\theta})$ 의 표준오차를 계산하기 위하여 델타(Delta) 방법(Bishop et al., 1975)을 이용한 점근적 표준오차(asymptotic standard error)는 다음의 식(1.1)으로 추정되며,

---

1) Associate Professor, Department of Statistics, Sungkyunkwan University.

E-mail: dkim@skku.ac.kr

2) Department of Statistics, Sungkyunkwan University.

$$\hat{\alpha}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11} + c} + \frac{1}{n_{12} + c} + \frac{1}{n_{21} + c} + \frac{1}{n_{22} + c}} \quad (1.1)$$

$\log \theta$ 에 관한  $100(1 - \alpha)\%$  신뢰구간은 식(1.2)와 같다.

$$\log(\hat{\theta}) \pm z_{\alpha/2} \hat{\alpha}(\log \hat{\theta}) \quad (1.2)$$

여기서,  $z_{\alpha/2}$ 는 표준정규분위수이며, 이 신뢰구간을 로짓 신뢰구간이라 한다. 일반적으로 대표본 근사방법은 소표본에서 정확하지 않다. Woolf(1955)와 Gart(1966)가 제안한 방법에서  $c=0$ 인 경우를 Woolf 방법이라 하며,  $c=0.5$ 인 경우를 Gart방법이라 하자. Gart방법은  $n_{ij}$ 가 0인 경우 발생하는 문제를 해결하였으나, 우리는 자료와 무관하게 각 칸에 0.5씩 더하는 것보다 더 나은 평활화(smoothing) 방법을 찾으려고 한다.

근사적 방법이 아닌 정확(exact) 신뢰구간은 그 포함확률이 명목수준보다 항상 크며 정확검정을 기초로 한 신뢰구간은 이산성에 기인된 보수성(conservativeness)이 발생한다. 이산형 자료의 보수성을 줄이기 위한 방법으로 Lancaster(1961)가 제안한 Mid-P방법이 있다. Mid-P방법은 이산형 변수의 이산성(discreteness)의 정도를 줄이며  $100(1 - \alpha)\%$  신뢰구간에서 포함확률이 정확검정(exact test)보다 명목수준에 더 근접하나 명목수준에 못 미칠 수 있다. Mid-P방법은 정확검정의 변형으로 검정 통계량의 정확분포(exact distribution)를 알아야 한다. 정규근사가 의심스러운 경우 변곡점 근사(saddlepoint approximation)방법은 분할표에서 정규근사보다 더 정확한 통계적 추론을 제공한다(Kim et al., 2000). 기존에 제시되었던 로짓 신뢰구간보다 좀더 향상된 방법을 찾기 위해 Agresti(1999)는  $2 \times 2$  분할표에서 칸도수를 독립적 모형방향으로 수정하는 방법을 제시하였다.

본 논문에서 2절은 유사 베이즈(pseudo-Bayes) 추정량에 대하여 살펴본다. 3절은  $2 \times 2$  분할표에서 유사 베이즈 추정량을 이용한 칸 도수 평활화 방법들을 제시하며, 이들 방법들을 사용하여 로그 오즈비에 대한 신뢰구간의 포함확률(coverage probability)과 평균길이(expected length)에 대하여 알아본다. 4절은 3절의 방법을  $2 \times 2 \times K$ 로 확장하여 분할표의 조건부 독립모형, 부분연관 모형 그리고 포화모형에서 추정된 칸 확률과 유사 베이즈 추정량을 이용하여 공통 오즈비(common odds ratio)에 대한 신뢰구간 추정방법들의 성능을 비교한다.

## 2. 유사 베이즈 추정량

$\mathbf{X} = \{n_1, \dots, n_l\}$ 가 모수  $N = \sum_{i=1}^l n_i$ ,  $\mathbf{p} = (p_1, \dots, p_l)$ 인 다항분포를 따른다고 하고  $\mathbf{x}$ 는  $\mathbf{X}$ 의 관찰된 벡터라 하자. 소표본인 경우  $\mathbf{X}$ 의 분포는 0이 많이 발생하므로  $\mathbf{X}$ 를 평활화하면 더 효율적인 추정값을 구할 수 있다. 우리는 원자료  $\mathbf{x}$ 나 표본비율  $\hat{\mathbf{p}} = \mathbf{x}/N$  대신 이들을 평활화 시킨 추정량을 로짓신뢰구간 구축에 사용하려 한다.  $\mathbf{p}$ 의 우도함수에 대한 사전분포(prior distribution)는 모수  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_l)$ 를 갖는 디리슈레(Dirichlet) 분포이며, 사후분포(posterior distribution)는 모수  $\boldsymbol{\beta} + \mathbf{x}$ 를 갖는 디리슈레 분포이다.  $W = \sum_{i=1}^l \beta_i$ ,  $\lambda_i = \beta_i/W$ 라 하면  $\mathbf{p}$ 의 사후평균인 베이즈 추정량  $\hat{\mathbf{q}}$ 는

$$\hat{q} = \hat{q}(W, \lambda) = \frac{N}{N+W} \cdot \frac{\mathbf{x}}{N} + \frac{W}{N+W} \cdot \lambda \quad (2.1)$$

이다.  $W$ 와  $\lambda$ 의 함수인  $\hat{q}$ 는  $\hat{p}$ 과  $\lambda$ 의 선형결합이며  $\hat{p}$ 를  $\hat{W}$ 의 크기에 따라  $\lambda$  방향으로 평활화한다.  $W$ 는 가중요소로 사용되며, 모수의 수  $t$ 가 클 때 일반적으로 베이즈 추정량은  $\hat{p}$ 보다 더 좋은 추정량이 된다.

만약  $\beta_i = 1/2$  이면,  $W$ 와  $\lambda$ 는 상수가 되며  $W = (1/2)t$ ,  $\lambda_i = 1/t$ 이 된다.  $p$ 의 사전평균인  $\lambda$ 에 대해  $\lambda_i = 1/t$ 은 각 칸에 동일확률을 부여하는 것이다. 즉 자료에 대한 정보가 없을 경우 동일 칸 확률을 사용하는 것은 하나의 대안이 될 수 있다. 그리고  $\lambda = (1/t)\mathbf{1}$ 는 모든  $\{n_i, i=1, \dots, t\}$ 가 서로 독립임을 의미한다. 이 경우  $p$ 의 베이즈 추정량은

$$\hat{q}(W, \lambda) = \frac{(n_1 + 1/2, \dots, n_t + 1/2)}{\sum_{i=1}^t n_i + \sum_{i=1}^t (1/2)} \quad (2.2)$$

이 된다. 따라서 동일 칸 확률  $\lambda_i = 1/t$  및  $W = (1/2)t$ 을 사용한 베이즈 추정량은 각 칸에서 칸도수에  $1/2$ 를 더하여 구한 상대도수가 된다. 이는 Gart방법과 일치한다. 따라서 Gart방법은  $\{n_i, i=1, \dots, n\}$ 가 독립이라는 독립성 모형하에서 칸 확률인  $\lambda = (1/t)\mathbf{1}$ 을 사용하고  $W = (1/2)t$ 를 사용한 유사 베이즈 추정량의 특별한 경우이다.  $2 \times 2$  분할표에서  $N = \sum \sum n_{ij}$ ,  $n_{i+} = \sum_j n_{ij}$ ,  $n_{+j} = \sum_i n_{ij}$ 라 할 경우 Agresti(1999)는  $\lambda_{ij} = 1/t$  대신 식(2.1)에  $\hat{\lambda}_{ij} = n_{i+} n_{+j} / N^2$ 를 사용하여  $n_{ij}$  대신  $n_{ij} + 2(n_{i+} n_{+j} / N^2)$ 로 칸 도수를 평활화 하였다.

베이즈 추정량  $\hat{q}(W, \lambda)$ 는  $W$ 와  $\lambda$ 에 의존하며  $\lambda$ 가  $W$ 에 영향을 미친다.  $W$ 를 추정하는 방법으로  $\lambda$ 와 자료와의 관계에 따라  $\lambda$ 가 자료에 독립으로 상수인 경우와  $\lambda$ 가 자료에 의존하는 경우로 나뉘어  $W$ 를 구할 수 있다. 이렇게  $W$ 와  $\lambda$ 를 추정한  $\hat{q}(\hat{W}, \hat{\lambda})$ 를 유사 베이즈(pseudo-Bayes) 추정량이라고 하며 다음과 같이 정의한다.

$$p^* = \hat{q}(\hat{W}, \hat{\lambda}) = \frac{N}{N+\hat{W}} \hat{p} + \frac{\hat{W}}{N+\hat{W}} \hat{\lambda} \quad (2.3)$$

$\lambda$ 가 자료에 독립일 경우  $\lambda$ 값은 총모수의 수에 역수를 취하여 사용한다(Fienberg와 Holland, 1972). 예를들어, 임의의 확률벡터  $p = \{p_1, \dots, p_t\}$ 에서 총모수의 수가  $t$ 일 때  $\lambda = (1/t)\mathbf{1}$ 를 사용한다.  $W$ 와  $\lambda$ 가 상수인 경우 추정량  $\hat{q}(W, \lambda)$ 와  $p$ 의 위험함수(risk function)를 구하여 위험함수를 최소화하는  $W$ 는

$$W = W(p, \lambda) = \frac{1 - \|p\|^2}{\|p - \lambda\|^2}$$

이다(Bishop et al., 1970).  $\mathbf{X}$ 가 독립인 다항분포를 따르는  $I \times J$ 분할표에서도  $\mathbf{p}$  대신  $\hat{\mathbf{p}}$ 을 사용하면  $\hat{W}$ 의 추정량은

$$\hat{W}^{(1)}_{I \times J} = \hat{W}^{(1)}_{I \times J}(\hat{\mathbf{p}}, \hat{\boldsymbol{\lambda}}) = \frac{N^2 - \sum_{i,j} n_{ij}^2}{\sum_{i,j} n_{ij}^2 - 2N \sum_{i,j} n_{ij} \hat{\lambda}_{ij} + N^2 \sum_{i,j} \hat{\lambda}_{ij}^2} \quad (2.4)$$

이다.

$\boldsymbol{\lambda}$ 를 추정하는 방법으로 다음의 세 가지를 고려할 수 있다. 첫째, 행과 열의 독립성 모형하에서의 칸 확률인  $\lambda_{ij} = n_{i+} n_{+j} / N^2$ 를 사용하여  $\hat{\mathbf{p}}$ 을 독립성 모형으로 평활화 시킨다. 둘째, 삼차원 분할표에서 예를 들면, 각 층의 행과 열의 관계(association)를 고려하여 반복비례적합 (iterative proportional fitting; IPF) 알고리즘을 사용하여  $\mathbf{x}$ 의 적합값(fitted value)인  $\hat{m}_{ijk}$ 를 구한후  $\hat{\lambda}_{ijk} = \hat{m}_{ijk} / N$ 을 사용한다. 여기서  $N = \sum \sum \sum n_{ijk}$ 이다. 셋째,  $\mathbf{X}$ 가 다항분포를 따를 때처럼  $\hat{\boldsymbol{\lambda}} = (1/I) \mathbf{1}$ 을 사용할 수 있으나 이차원 분할표에서 동일 확률의 가정은 적절하지 않다.  $\boldsymbol{\lambda}$ 의 추정값  $\hat{\boldsymbol{\lambda}}$ 이 정해지면  $\hat{W}$ 이 결정되어 유사 베이즈 추정량  $\mathbf{p}^*$ 가 구해진다.

$\mathbf{X} = (\{n_{ij}\})$ 가  $I \times J$ 분할표인 경우  $\lambda_{ij} = n_{i+} n_{+j} / N^2$ 을 사용하여 Sutherland(1974)는  $\hat{W}$ 를 다음과 같이 추정하였다(Bishop et al., 1970, p432).

$$\hat{W}^{(2)}_{I \times J} = \frac{A_N + B_N Q_0 - C_N(Q_1 + Q_2 + Q_3)}{D_N Q_3 - E_N Q_0 + F_N Q_1 Q_2 + G_N(Q_1 + Q_2) - (N-1)} \quad (2.5)$$

여기서,  $Q_0 = \sum_{i,j} n_{ij} n_{i+} n_{+j}$ ,  $Q_1 = \sum_i n_{i+}^2$ ,  $Q_2 = \sum_j n_{+j}^2$ ,  $Q_3 = \sum_{i,j} n_{ij}^2$ ,  $A_N = N(N-1)$ ,  $B_N = \frac{2N-1}{N^2}$ ,

$$C_N = \frac{N-1}{N}, \quad D_N = \frac{N^5 - 5N^4 + 7N^3 - 3N^2 + 24}{N^4(N-3)}, \quad E_N = \frac{2(N-1)^2}{N^3}, \quad F_N = \frac{(N-1)(N-2)}{N^4},$$

$$G_N = \frac{N^3 - 12N^2 + 17N - 8}{N^3(N-3)}$$

이다.

### 3. 2×2 분할표에서 로짓 신뢰구간

2×2 분할표에서  $\mathbf{X} = (\{n_{ij}\})$ 를 칸도수라 하며  $N = \sum \sum n_{ij}$ 라 하자. 우리는 소표본일 경우 Woolf, Gart, Agresti의 칸 수정방법과 유사 베이즈 추정량을 이용한 칸도수 평활화 방법을 로그 오즈비 신뢰구간 추정에 적용한다. 그리고 각 방법들의 성능 비교를 위해 우리는 모의실험을 통하여 여러 방법의 신뢰구간의 포함확률과 평균길이를 구한다.

모의실험에서 행(row)합계를 고정시킨 후  $n_{11} \sim B(n_1, p_1)$  그리고  $n_{21} \sim B(n_2, p_2)$ 인 독립적인 이항(binomial)난수를 발생하여 2×2 분할표를 생성시킨 후 식(1.2)의  $\log \theta$ 에 대한 신뢰구간을 계산한다. 이와같은 과정을 10,000번 시행하여 10,000개의 신뢰구간 중 신뢰구간이 모수  $\log \theta$ 를 포함하는 포함확률을 구한다. 그리고 2×2 분할표에서 신뢰구간의 평균길이는 각 분할표에서 생성된

이항난수  $n_{ij}$ 가 0보다 큰 경우만 계산하여 여러 방법들의 결과를 비교한다.

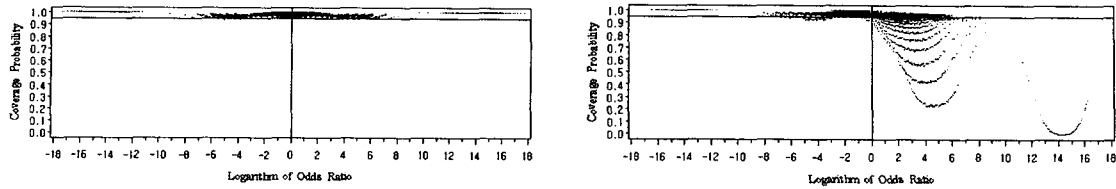
### 3.1 가중요소 $W$ 이 상수인 경우

다항확률  $p$ 의 베이즈 추정량 식(2.3)에서  $W=(1/2)t=2$ 로 하면 칸수정값은  $n_{ij}^* = n_{ij} + 2\lambda_{ij}$ 이 되며 평활화된  $n_{ij}^*$ 를 사용하여 로짓 신뢰구간을 만든다.  $\lambda_{ij}$ 의 추정량으로  $\lambda_{ij}=0, 1/4, n_{i+} \cdot n_{+j}/N^2$ 를 사용하면 이들 방법이 각각 Woolf, Gart, Agresti방법이다. Woolf방법은 각 칸에 수정을 하지 않고 원자료를 이용하므로 분할표의 한 칸이 0을 갖는 경우 로그오즈비 또는 표준오차가 존재하지 않는다. 따라서 로짓 신뢰구간이 정의되지 않는 경우 신뢰구간을  $(-\infty, \infty)$ 으로 정의하기로 한다. Gart방법은 각 칸에  $c=0.5$ 를 더하므로 신뢰구간이 존재하지 않는 분할표는 발생하지 않는다. 독립 모형으로 평활화 시키는 방법에서  $n_{11}=n_{21}=0$  또는  $n_{12}=n_{22}=0$ 인 경우 신뢰구간이 존재하지 않으므로 신뢰구간을  $(-\infty, \infty)$ 으로 정의하기로 한다. <표 3.1>은  $2 \times 2$  분할표에서 각 행의 표본수가  $n_1=n_2=10$ 인 경우와  $n_1=30, n_2=10$ 인 경우에 모수  $p_1$ 과  $p_2$ 를 따르는 이항랜덤표본을 이용하여 구한 로짓 신뢰구간의 포함확률과 평균길이를 나타낸다.

<그림 3.1>에서 <그림 3.3>은 세 방법의 포함확률을 나타낸다.  $p_1$ 과  $p_2$ 는 각각 0.01부터 1.0까지 0.01씩 증가하게 하여  $p_1$ 과  $p_2$ 에 따른 10,000개의 포함확률을 그린 것이다. Woolf방법을 사용한 경우  $n_1=n_2=10$ 인 경우 로그 오즈비가 커질수록 포함확률이 명목신뢰수준보다 크다는 것을 알 수 있다. 그러나 Woolf방법에서 각 행의 표본수가 다른 경우에는 로그 오즈비가 양의 방향으로 커질 경우에는 포함확률이 낮아지는 것을 볼 수 있다. Gart방법은 Woolf방법보다 포함확률이 명목수준에 가깝고 신뢰구간의 길이도 줄어들었으나 로그 오즈비가 큰 값일 경우 포함확률이 좋지 않게 나타났다. Agresti(1999)의 독립적 평활화 방법은 Gart방법과 포함확률이 비슷하나 신뢰구간의 평균길이는 Gart방법보다 약간 크고 Woolf방법보다 작다. Agresti방법에서도 신뢰구간이 존재하지 않는 경우 신뢰구간 수정을 하지만 Woolf방법의 경우보다는 작으며, Agresti방법이 Gart방법보다 포함확률이 명목신뢰수준에 못 미치는 경우가 더 작다.

<표 3.1> 로짓 신뢰구간의 포함확률과 평균길이( $n_{ij}^* = n_{ij} + 2\lambda_{ij}$ )

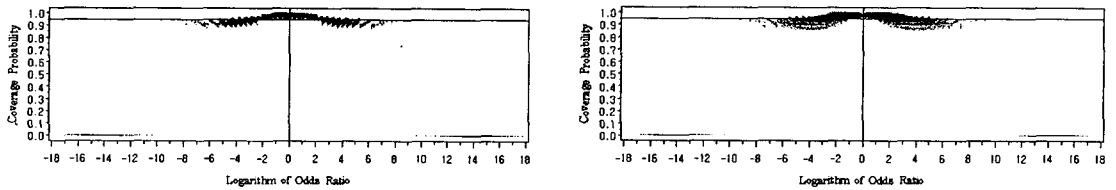
		$n_1 = n_2 = 10$						$n_1 = 30, n_2 = 10$					
		Woolf		Gart		Agresti		Woolf		Gart		Agresti	
$p_1$	$p_2$	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이
0.5	0.5	0.9603	3.74	0.9583	3.51	0.9583	3.54	0.9659	3.02	0.9672	2.87	0.9658	2.93
0.5	0.1	0.9736	4.54	0.9619	4.03	0.9627	4.16	0.9732	3.95	0.9565	3.47	0.9614	3.71
0.4	0.2	0.9841	4.25	0.9806	3.85	0.9823	3.98	0.9753	3.57	0.9652	3.24	0.9753	3.43
0.4	0.1	0.9747	4.58	0.9731	4.06	0.9745	4.25	0.9728	3.97	0.9596	3.49	0.9652	3.76
0.3	0.3	0.9800	4.14	0.9779	3.79	0.9779	3.91	0.9730	3.34	0.9740	3.10	0.9730	3.24
0.3	0.1	0.9886	4.71	0.9813	4.15	0.9882	4.41	0.9733	4.02	0.9610	3.53	0.9733	3.85
0.2	0.05	0.9946	5.10	0.9853	4.40	0.9944	4.81	0.9724	4.36	0.9418	3.78	0.9718	4.21
0.1	0.1	0.9999	5.22	0.9998	4.48	0.9998	4.94	0.9841	4.54	0.9808	3.96	0.9841	4.42
0.1	0.05	0.9999	5.38	0.9997	4.58	0.9999	5.10	0.9809	4.73	0.9627	4.08	0.9809	4.61



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

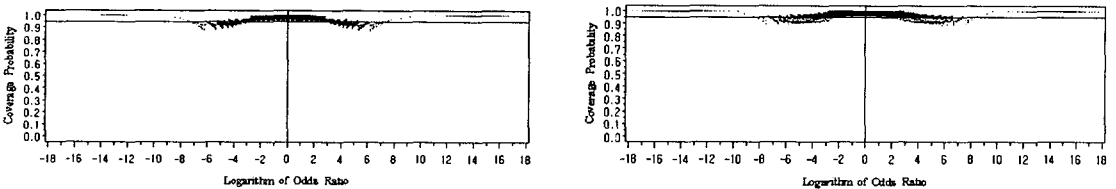
<그림 3.1> Woolf 방법의 로짓 신뢰구간의 포함확률



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

<그림 3.2> Gart 방법의 로짓 신뢰구간의 포함확률



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

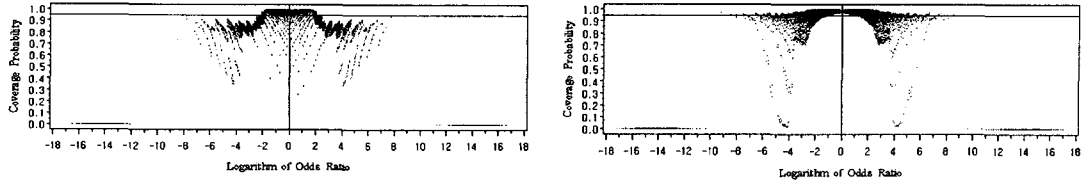
<그림 3.3> Agresti 방법의 로짓 신뢰구간의 포함확률

3.2 위험함수를 고려한 가중요소  $\hat{W}$

2장에서 설명한  $p$ 의 유사 베이지 추정량  $p^*$ 를 이용하여 로그 오즈비의 포함확률과 평균길이를 비교하고자 한다. 3.1절에서 가중요소  $\hat{W}$ 에 상수를 사용하는 대신 자료의 정보를 이용하여 위험 함수를 최소화하는  $W$ 의 추정량인 식(2.4) 또는 식(2.5)를 사용한다. 그리고  $\lambda$ 의 추정량으로 동일 확률  $1/4$  또는  $n_{i+}n_{+j}/N^2$ 를 사용하여  $p$ 의 유사 베이지 추정량 식(2.3)을 구한다. 유사 베이지 추정량  $p_{ij}^*$ 를 이용한 평활화된 칸 추정값  $n_{ij}^*$ 는

$$n_{ij}^* = N \cdot p_{ij}^* \tag{3.1}$$

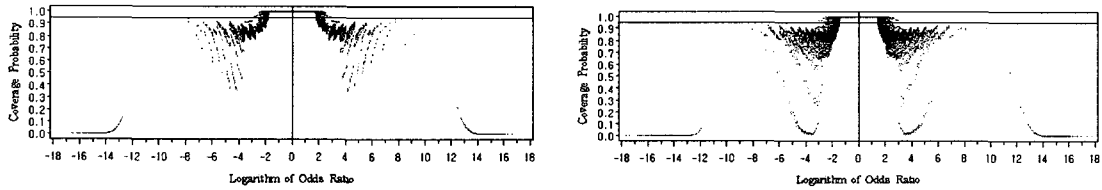
이다. 유사 베이지 추정량에서  $\hat{W}$ 를 계산할 때  $\hat{\theta}=1$ 인 경우  $\hat{W}$ 의 분모가 0이 된다. 본 논문에서는 이런 경우에는  $\hat{W}$ 의 분모에는 0.1로 수정했으며,  $N \cdot p_{ij}^*=0$ 인 경우에는 오즈비 추정량과  $\hat{\alpha}(\log \hat{\theta})$ 가 존재하지 않기 때문에 신뢰구간을  $(-\infty, \infty)$ 으로 정의하였다. <그림 3.4>부터 <그림 3.6>은 평활화된 칸 추정값  $N \cdot p_{ij}^*$ 를 사용하여 구한 포함확률을 나타낸다.  $\hat{\lambda}_{ij}=n_{i+}n_{+j}/N^2$ 를 사용한 경우 오즈비가  $e^{-2}=0.14$ 에서  $e^2=7.39$  사이에 있으면 포함확률이 거의 1.0으로 나타났다. 그러나 오즈비가 이 구간 밖에 있는 경우 포함확률이 낮아진다.  $\hat{W}$ 값은 일반적으로 2 이상이며



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

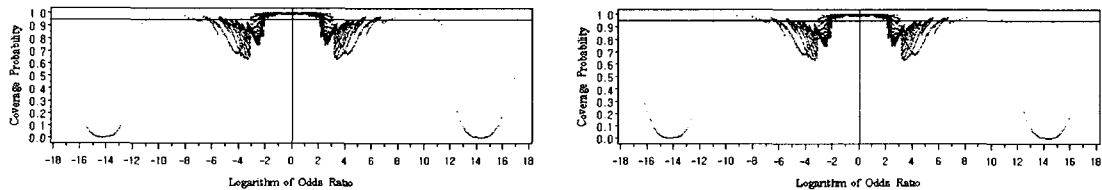
<그림 3.4> 평활화된 칸 추정값  $n_{ij}^* = N\hat{p}_{ij}^*$ 를 이용한 로짓 신뢰구간의 포함확률 ( $\hat{W}$ =식(2.4),  $\lambda_{ij} = 1/4$ )



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

<그림 3.5> 평활화된 칸 추정값  $n_{ij}^* = N\hat{p}_{ij}^*$ 를 이용한 로짓 신뢰구간의 포함확률 ( $\hat{W}$ =식(2.4),  $\lambda_{ij} = n_{i+} n_{+j}/N^2$ )



a)  $n_1 = n_2 = 10$

b)  $n_1 = 30, n_2 = 10$

<그림 3.6> 평활화된 칸 추정값  $n_{ij}^* = N\hat{p}_{ij}^*$ 를 이용한 로짓 신뢰구간의 포함확률 ( $\hat{W}$ =식(2.5),  $\lambda_{ij} = n_{i+} n_{+j}/N^2$ )

매우 큰 경우는 자료를  $\hat{\lambda}$ 의 방향으로 많이 변화시킨다. 이 방법은  $2 \times 2$  분할표에서 포함확률이 좋지 않아 포함확률에 관한 표는 생략하였다.

### 3.3 유사 베이즈 추정량 $\hat{p}^*$ 방향으로 평활화

$p$ 의 베이즈 추정량 식(2.3)에서 우리는 가중요인  $\hat{W}$ 에 대한 추정량으로  $(1/2)t$  또는 식(2.4)나 식(2.5)를 사용할 수 있다. Gart나 Agresti가  $\hat{p} = \mathbf{x}/N$ 의 평활화를 위해  $\hat{\lambda}_{ij}$ 에  $1/t$  또는  $n_{i+} n_{+j}/N^2$ 를 사용하며  $n_{ij} + 2\hat{\lambda}_{ij}$ 으로 칸 수정을 하였다. 우리는 자료의 정보를 최대한 활용해서 평활화하기 위해  $\hat{\lambda}$ 에 베이즈 추정량  $\hat{p}^*$ 를 사용한다. 따라서  $n_{ij}$ 를 평활화한

$$n_{ij}^* = n_{ij} + 2\hat{p}_{ij}^* \quad (3.2)$$

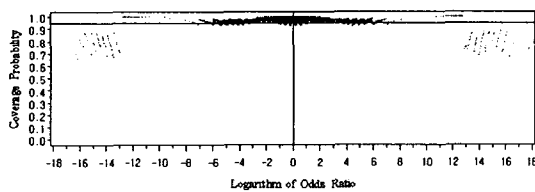
를 사용하여 로짓 신뢰구간을 구축한다. 3.2절에서와 같이  $\hat{W}$ 과  $\hat{\lambda}$ 에 따른 식(2.4)와 식(2.5)의 계산시  $\hat{W}$ 의 분모가 0인 경우 0.1로 수정하며, 신뢰구간이 존재하지 않는  $n_{11} = n_{21} = 0$  과

<표 3.2> 로짓 신뢰구간의 포함확률(  $n_{ij}^* = n_{ij} + 2p_{ij}^*$  )

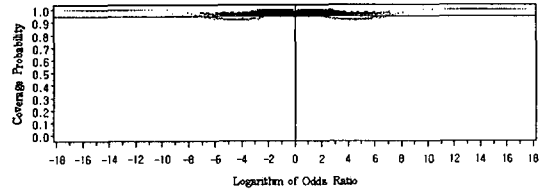
$p_1$	$p_2$	$n_1 = n_2 = 10$					$n_1 = 30, n_2 = 10$				
		$\hat{W} = (1/2)t$		$\hat{W}^{(1)}_{I \times J}$		$\hat{W}^{(2)}_{I \times J}$	$\hat{W} = (1/2)t$		$\hat{W}^{(1)}_{I \times J}$		$\hat{W}^{(2)}_{I \times J}$
		$\lambda_{ij} = \frac{1}{4}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$	$\lambda_{ij} = \frac{1}{4}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$	$\lambda_{ij} = \frac{1}{4}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$	$\lambda_{ij} = \frac{1}{4}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$	$\lambda_{ij} = \frac{n_i+n_j}{N^2}$
0.5	0.5	0.9603	0.9603	0.9603	0.9599	0.9603	0.9601	0.9601	0.9601	0.9668	0.9601
0.5	0.1	0.9736	0.9736	0.9628	0.9598	0.9628	0.9732	0.9732	0.9648	0.9628	0.9648
0.4	0.2	0.9692	0.9692	0.9692	0.9763	0.9692	0.9753	0.9753	0.9753	0.9743	0.9753
0.4	0.1	0.9747	0.9747	0.9747	0.9733	0.9746	0.9728	0.9728	0.9653	0.9621	0.9728
0.3	0.3	0.9800	0.9800	0.9800	0.9835	0.9800	0.9722	0.9722	0.9722	0.9720	0.9722
0.3	0.1	0.9855	0.9855	0.9855	0.9869	0.9855	0.9695	0.9695	0.9695	0.9737	0.9695
0.2	0.05	0.9946	0.9946	0.9946	0.9941	0.9946	0.9599	0.9599	0.9599	0.9695	0.9718
0.1	0.1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9841	0.9841	0.9841	0.9834	0.9841
0.1	0.05	0.9999	0.9999	0.9999	0.9991	0.9999	0.9809	0.9809	0.9809	0.9814	0.9809

$n_{12} = n_{22} = 0$  경우 신뢰구간은  $(-\infty, \infty)$  으로 정의한다. <표 3.2>와 <그림 3.7> 부터 <그림 3.11> 은  $\hat{W}$ 과  $\lambda$ 에 따른 5가지 경우의 유사 베이즈 추정량  $p_{ij}^*$ 를 구하여  $n_{ij} = n_{ij} + 2p_{ij}^*$ 를 이용한 로그 오즈비에 대한 신뢰구간의 포함확률을 나타낸다.

3.1절의  $\lambda_{ij}$  대신 유사 베이즈 추정량  $p_{ij}^*$ 를 사용한  $n_{ij} + 2p_{ij}^*$  방법은 포함확률이 0.95이하인 경우가 많이 줄었다. 특히  $\hat{W} = (1/2)t$  이고  $\lambda_{ij} = n_i + n_j / N^2$ 를 사용하여 유사 베이즈 추정량  $p_{ij}^*$ 를 구한 후  $n_{ij}^* = n_{ij} + 2p_{ij}^*$ 로 평활화 시킨 방법이 3.1절의 Woolf, Gart나 Agresti방법보다 포함확률이 명목수준에 못미치는 경우가 작게 나타났으며, 신뢰구간의 평균길이는 Agresti 방법보다 약간 크게 나타났다. 그러나 위험함수를 최소화하는  $\hat{W}^{(1)}_{I \times J}$ ,  $\hat{W}^{(2)}_{I \times J}$ 를 사용한 경우 로그 오즈비의 절댓값이 11이상인 경우는 명목수준보다 작아지다 증가하는 경향이 있어  $\hat{W} = (1/2)t$  대신 위험함수를 최소화하는  $\hat{W}$ 를 사용하는 방법이 더 나은 결과를 제공하지 않았다. 따라서  $2 \times 2$  분할표에서  $\hat{W} = (1/2)t$ ,  $\lambda_{ij} = n_i + n_j / N^2$ 를 사용하여  $n_{ij}^* = n_{ij} + 2p_{ij}^*$ 로 평활화 시키는 것이 포함확률 측면에서 가장 좋게 나타났다. 그리고 주어진 표본 크기에서 <표 3.2>의  $\hat{W}$ 과  $\lambda$ 에 따른 각 방법에 따라 평균길이는 차이가 나타나지 않았으며 표본수가 늘수록 평균길이는 줄어들었다.



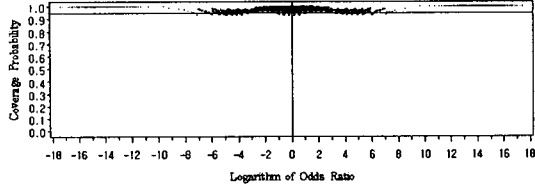
a)  $n_1 = n_2 = 10$



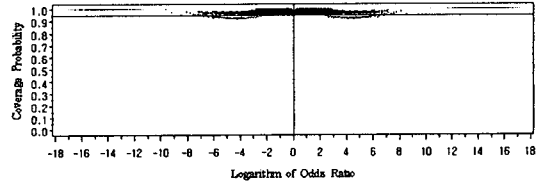
b)  $n_1 = 30, n_2 = 10$

<그림 3.7>  $\hat{W} = (1/2)t$ ,  $\lambda_{ij} = 1/4$  인 경우의 로짓 신뢰구간의 포함확률



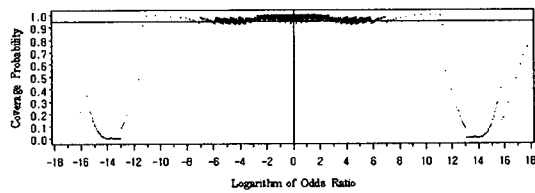


a)  $n_1 = n_2 = 10$

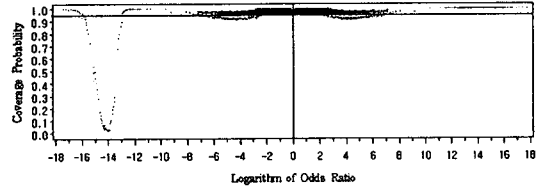


b)  $n_1 = 30, n_2 = 10$

<그림 3.8>  $\hat{W} = (1/2)t$ ,  $\lambda_{ij} = n_{i+}n_{+j}/N^2$  인 경우의 로짓 신뢰구간의 포함확률

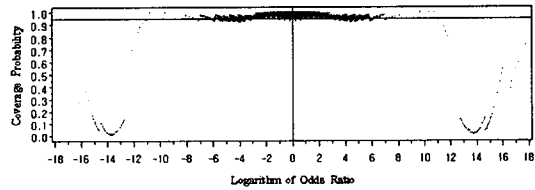


a)  $n_1 = n_2 = 10$

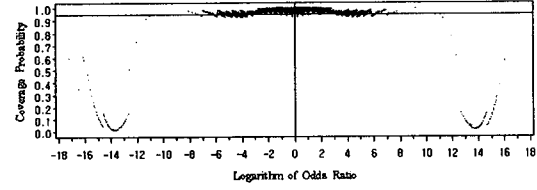


b)  $n_1 = 30, n_2 = 10$

<그림 3.9>  $\hat{W}$  = 식(2.4),  $\lambda_{ij} = 1/4$  인 경우의 로짓 신뢰구간의 포함확률

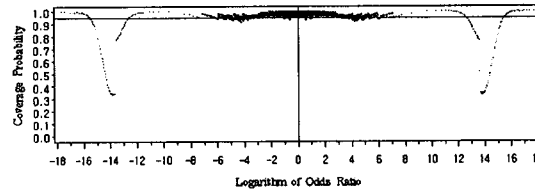


a)  $n_1 = n_2 = 10$

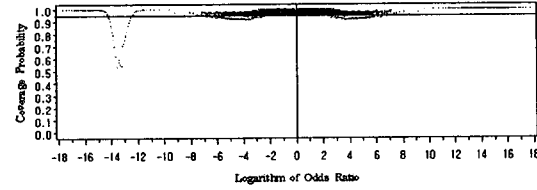


b)  $n_1 = 30, n_2 = 10$

<그림 3.10>  $\hat{W}$  = 식(2.4),  $\lambda_{ij} = n_{i+}n_{+j}/N^2$  인 경우의 로짓 신뢰구간의 포함확률



a)  $n_1 = n_2 = 10$



b)  $n_1 = 30, n_2 = 10$

<그림 3.11>  $\hat{W}$  = 식(2.5),  $\lambda_{ij} = n_{i+}n_{+j}/N^2$  인 경우의 로짓 신뢰구간의 포함확률

#### 4. $2 \times 2 \times K$ 분할표에서의 로짓 신뢰구간

$2 \times 2 \times K$  분할표에서  $\mathbf{X} = (\{n_{ijk}\})$  를 칸도수라 하며  $N = \sum \sum \sum n_{ijk}$  라 하자. 우리는  $2 \times 2 \times K$  분할표에서 유사 베이즈 추정량을 이용하여 공통 오즈비(common odds ratio) 신뢰구간의 추정에 대하여 알아보기로 한다.  $2 \times 2 \times K$  분할표에서 공통 오즈비  $\theta_{11(1)} = \dots = \theta_{11(K)}$  의 추정량은 식(4.1)과 같다(Mantel과 Haenszel, 1959).

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{\sum_k (n_{12k} n_{21k} / n_{++k})} \tag{4.1}$$

$K=1$ 이면  $\hat{\theta}_{MH}$ 는  $2 \times 2$ 분할표에서의 오즈비와 같다. 또한,  $2 \times 2 \times K$  분할표에서  $\log(\hat{\theta}_{MH})$ 의 분산 추정량은 식(4.2)와 같다(Robins et al., 1986).

$$\begin{aligned} \hat{\sigma}^2[\log(\hat{\theta}_{MH})] = & \frac{\sum (n_{11k} + n_{22k})(n_{11k} n_{22k}) / n_{++k}^2}{2(\sum n_{11k} n_{22k} / n_{++k})^2} + \frac{\sum (n_{12k} + n_{21k})(n_{12k} n_{21k}) / n_{++k}^2}{2(\sum n_{12k} n_{21k} / n_{++k})^2} \\ & + \frac{\sum [(n_{11k} + n_{22k})(n_{12k} n_{21k}) + (n_{12k} + n_{21k})(n_{11k} n_{22k})] / n_{++k}^2}{2(\sum n_{11k} n_{22k} / n_{++k})(\sum n_{12k} n_{21k} / n_{++k})} \end{aligned} \tag{4.2}$$

3절에서는  $2 \times 2$  분할표에서 유사 베이즈 추정량  $p_{ij}^*$ 를 사용하여  $n_{ij}$ 를 평활화 시켰다. 4절에서는  $2 \times 2 \times K$  분할표로 칸도수의 평활화 방법을 확장한다.  $2 \times 2 \times K$  분할표의 각 층에서 이항랜덤표본의 확률모수를  $p_{1k}$ 와  $p_{2k}$ 라 하면 공통 오즈비는 다음과 같이 정의한다.

$$\theta_{MH} = \frac{\sum_k (n_{1k} p_{1k})(n_{2k} - n_{2k} p_{2k}) / n_{++k}}{\sum_k (n_{1k} - n_{1k} p_{1k})(n_{2k} p_{2k}) / n_{++k}} \tag{4.3}$$

$2 \times 2 \times K$ 분할표에서 공통 오즈비의 신뢰구간 추정을 위해 각 층에서 오즈비가 가능한 비슷하도록  $p_{1k}$ 와  $p_{2k}$ 를 정해야 한다. 각 층의 오즈비가 같다는 귀무가설 하에서 각 층의 오즈비  $\theta_{11(k)}$ 와 두 번째 행에 대한 모수  $p_{2k}$ 가 주어지면 첫 번째 행 모수  $p_{1k}$ 는 다음과 같이 결정된다.

$$p_{1k} = \left( \frac{1 - p_{2k}}{1 - p_{2k} + p_{2k} \theta_{11(k)}} \right) \left( \frac{p_{2k}}{1 - p_{2k}} \right) \theta_{11(k)} \tag{4.4}$$

$2 \times 2 \times K$ 분할표에서 공통 오즈비에 관한 신뢰구간 추정방법의 성능비교를 위해 모의실험에서 이항난수를 발생시킬 때 각 층의 연관정도가 많이 바뀌면 공통 오즈비 통계량의 성능은 매우 좋지 않게 된다. 따라서 각 층에서 두 번째 행 모수  $p_{2k}$ ,  $k=1, 2$ , 가 주어진 경우 각 층의 오즈비가 크게 변하지 않게 이항난수를 발생시켜 각 신뢰구간이 공통 오즈비를 포함하는 포함확률과 평균 길이를 구한다.  $2 \times 2 \times K$  분할표에서 칸 수정을 하지 않는 방법과 칸에 동일한 수정값 0.5를 더하는 Gart방법과 로그선형모형에서 추정된 칸 확률을 이용한 평활화 방법을 고려한다.

3절의  $2 \times 2$  분할표에서처럼  $2 \times 2 \times K$  분할표에서 유사 베이즈 추정량  $p^*$ 는

$$p^* = \frac{N}{N + \hat{W}} \hat{p} + \frac{\hat{W}}{N + \hat{W}} \hat{\lambda} \tag{4.5}$$

이다.  $t$ 를 다항확률  $p$ 의 개수라 할때  $\lambda$ 의 추정값으로  $p$ 의 동일확률  $1/t$ 뿐 아니라 로그선형

모형에서 조건부 독립(conditional independence)모형, 부분연관(no three-factor interaction)모형, 그리고 포화(saturated)모형의 칸 확률을 사용할 수 있다. 따라서  $\lambda$ 의 추정량으로

$$\lambda_{ijk} = \begin{cases} 1/t & , \quad \text{동일 확률 모형} \\ \frac{n_{i+k} n_{+jk}}{n_{++k}} \cdot \frac{1}{N} & , \quad \text{조건부 독립 모형}(XZ, YZ) \\ \frac{\hat{m}_{ijk}(\text{IPF})}{N} & , \quad \text{부분 연관 모형}(XY, XZ, YZ) \\ \frac{n_{ijk}}{N} & , \quad \text{포화 모형}(XYZ) \end{cases} \quad (4.6)$$

를 사용한다. 부분 연관 모형에서는 칸의 적합값이 일정한 형태로 정해지지 않으므로 반복비례적합 알고리즘을 사용하여 칸 적합값 및 칸 확률을 구한다. 우리는 소표본이며 층의 개수가 작은 경우 어떤 방법의 로짓 신뢰구간이 명목신뢰수준을 더 잘 만족하는지를 연구한다.

#### 4.1 가중요소 $\hat{W}$ 이 상수인 경우

$2 \times 2 \times 2$  분할표에서 유사 베이지 추정량  $p^*$  계산시  $\hat{W} = (1/2)t = 4$ 를 사용하면  $p^*$ 는  $p^* = \frac{N}{N+4} \left( \frac{n_{ijk}}{N} \right) + \frac{4}{N+4} \hat{\lambda}$ 이며 평활화된 칸 추정값  $n_{ijk}^*$ 는

$$n_{ijk}^* = n_{ijk} + 4 \hat{\lambda}_{ijk} \quad (4.7)$$

이 된다.  $K=2$ 를 고려한 이유는  $2 \times 2$  분할표에서처럼 칸 도수에 0이 발생할 경우 공통오즈비나  $\sigma^2[\log(\hat{\theta}_{MH})]$  추정에 영향을 많이 미치기 때문이다.

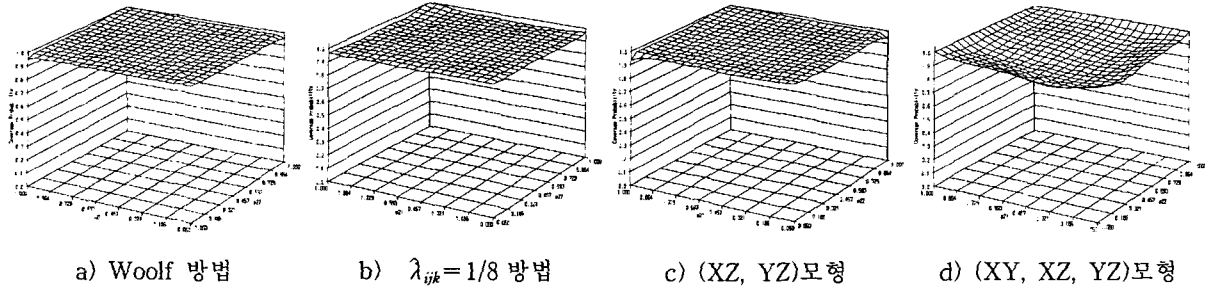
소표본으로 행 표본수가  $n_{1k} = n_{2k} = 10$ 인 경우 Woolf방법,  $\hat{\lambda}_{ijk} = 1/8$ 인 Gart방법과 세 가지 로그선형모형에서 추정된 칸 확률을 이용하여 로그오즈비의 신뢰구간에 대한 포함확률과 평균길이를 비교한다. Woolf 방법, (XZ, YZ)모형, (XY, XZ, YZ)모형, (XYZ)모형을 이용하여 칸 수정을 할 경우 신뢰구간이 정의되지 않는 경우가 발생한다. Woolf방법과 (XYZ)모형을 이용할 때는 로그 오즈비가 존재하지 않는  $n_{121} = n_{212} = 0$ ,  $n_{211} = n_{122} = 0$  경우와 표준오차가 존재하지 않는  $n_{i11} = n_{i22} = 0$  (단,  $i, j = 1, 2$ )인 경우, 그리고 (XZ, YZ)모형을 이용할 때는 표준오차가 존재하지 않는  $n_{111} = n_{211} = n_{112} = n_{212} = 0$  또는  $n_{121} = n_{221} = n_{122} = n_{222} = 0$ 인 경우, 그리고 (XY, XZ, YZ)모형을 이용할 때는 반복비례적합값이 존재하지 않는  $n_{11k} = n_{21k} = 0$  또는  $n_{12k} = n_{22k} = 0$  또는  $n_{i11} = n_{i22} = 0$  (단,  $i, j, k = 1, 2$ )인 경우에 대하여 신뢰구간이 정의되지 않으므로 3절에서와 같이 신뢰구간을  $(-\infty, \infty)$ 으로 정의한다. 단,  $\hat{\lambda}$ 에 동일확률을 주는 Gart방법은 이와 같은 수정을 하지 않는다.

<표 4.1>은  $\theta_{11(k)} = 5$ 인 경우에  $n_{ijk} + 4 \hat{\lambda}_{ijk}$ 을 사용하여 신뢰구간의 포함확률과 평균길이를 나타낸다. 주어진  $p_{21}$ 과  $p_{22}$ 에서 모의실험의 시행횟수는 10,000번이다.  $p_{21}$ 과  $p_{22}$ 를 0.05부터 1.0까지

<표 4.1>  $\theta_{11(k)}=5$  경우 로그 공통 오즈비에 대한 포함확률과 평균길이( $n_{1k}=n_{2k}=10, k=1,2$ )

$$n_{ijk}^* = n_{ijk} + 4 \lambda_{ijk}$$

$p_{21}$	$p_{22}$	Woolf 방법		$\lambda_{ijk}=1/8$		(XZ, YZ)		(XY, XZ, YZ)		(XYZ)	
		포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이
0.1	0.1	0.9749	3.16	0.9596	2.83	0.9700	2.96	0.9294	2.44	0.9684	3.01
0.2	0.1	0.9745	3.00	0.9636	2.72	0.9666	2.81	0.8978	2.29	0.9666	2.86
0.3	0.2	0.9691	2.84	0.9604	2.62	0.9615	2.65	0.8674	2.15	0.9593	2.71
0.4	0.3	0.9659	2.82	0.9573	2.61	0.9578	2.63	0.8647	2.13	0.9526	2.69
0.5	0.5	0.9715	2.95	0.9611	2.69	0.9634	2.75	0.8856	2.24	0.9623	2.81
0.6	0.2	0.9717	2.97	0.9599	2.70	0.9629	2.77	0.8889	2.26	0.9614	2.83
0.7	0.3	0.9745	3.00	0.9601	2.72	0.9647	2.81	0.8872	2.28	0.9648	2.86
0.8	0.7	0.9798	3.38	0.9587	2.98	0.9788	3.19	0.9517	2.65	0.9757	3.23
0.9	0.9	0.9941	3.79	0.9687	3.24	0.9938	3.60	0.9794	3.06	0.9898	3.61



<그림 4.1>  $\theta_{11(k)}=5$ 인 경우 로그 공통오즈비에 대한 각 방법별 포함확률( $n_{ijk}^* = n_{ijk} + 4 \lambda_{ijk}$ )

0.05씩 증가시켜  $p_{21}$ 과  $p_{22}$ 에 따른 400개의 포함확률의 결과를 <그림 4.1>에 나타냈다.

$\theta_{11(k)}$ 가 변화함에 따라서 Woolf방법의 포함확률이 변화하지만 명목신뢰수준보다 크게 나타난다. Gart방법의 포함확률은  $\theta_{11(k)}=5$ 를 기준으로 5보다 작을 경우에는 포함확률이 명목신뢰수준보다 크게 나타나고 5인 경우에는 포함확률이 명목신뢰수준에 가까운 값을 가지며  $\theta_{11(k)}$ 가 5보다 커질수록 포함확률이 명목신뢰수준보다 작아지면서 포함확률이 좋지 않게 나타난다. Woolf방법은 신뢰구간이 존재하지 않아서 신뢰구간을 수정한 경우가 많다.

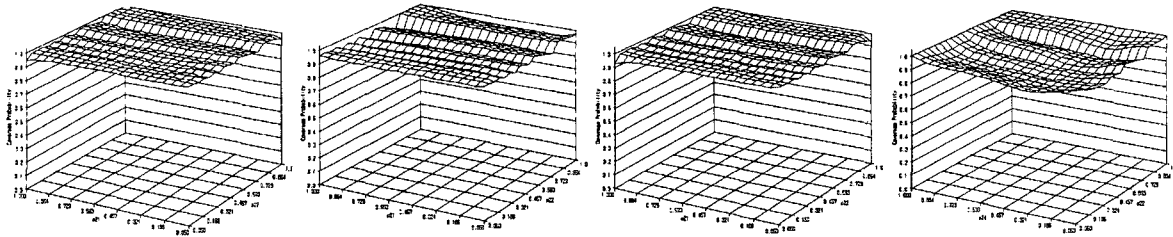
<표 4.1>에서  $\lambda$ 의 추정값으로 (XZ, YZ)모형의 칸 확률을 사용한 포함확률은 대체로 (XYZ)모형과 Woolf방법의 포함확률 사이에 있으며 포함확률의 그림도 이들 방법과 비슷하게 나타났다. 신뢰구간의 평균길이에 대한 결과를 비교하면 (XY, XZ, YZ)모형을 이용한 방법의 평균길이가 가장 짧게 나타났으나 포함확률은 좋지 않다. 그러므로 위의 방법들 중에서 포함확률이 대체로 명목수준을 만족하며 신뢰구간의 수정을 최소로하는 조건부 독립 모형 (XZ, YZ) 방법이 가장 좋다.

다음은 두 번째 층의 오즈비가 5에서 변동하게 하여 공통 오즈비에 대한 신뢰구간 추정방법의 성능을 비교한다. 우리는  $\theta_{11(k)}$ 를 고정하고  $p_{22}$ 가 주어질 경우 계산되어지는  $p_{12}$ 의 값을 반올림하여 자료를 생성한다. 따라서 첫 번째 층은 오즈비가 5를 만족하나 두 번째 층의 오즈비가 3~7 사이로 변화하여 공통 오즈비는  $5 \pm 0.2$  만큼 변한다. <표 4.2>와 <그림 4.2>는  $p_{12}$ 를 반올림하여 공통 오즈비가 5에서 변할 경우 포함확률을 나타낸다.

<표 4.2>  $\theta_{MH}$  가 5에서 변할 경우 로그 오즈비에 대한 포함확률과 평균길이( $n_{1k} = n_{2k} = 10, k = 1, 2$ )

$$n_{ijk}^* = n_{ijk} + 4 \lambda_{ijk}$$

$p_{21}$	$p_{22}$	Woolf		$\lambda_{ijk} = 1/8$		(XZ, YZ)		(XY, XZ, YZ)		(XYZ)	
		포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이
0.1	0.1	0.9784	3.63	0.9620	3.04	0.9721	3.29	0.9348	2.69	0.9733	3.46
0.2	0.1	0.9759	3.35	0.9639	2.89	0.9688	3.04	0.9002	2.49	0.9685	3.19
0.3	0.2	0.9695	2.97	0.9648	2.69	0.9662	2.73	0.8642	2.21	0.9561	2.84
0.4	0.3	0.9653	2.92	0.9595	2.66	0.9610	2.70	0.8685	2.18	0.9520	2.79
0.5	0.5	0.9575	3.10	0.9361	2.75	0.9444	2.85	0.8726	2.32	0.9457	2.95
0.6	0.2	0.9763	3.30	0.9639	2.86	0.9704	2.99	0.8917	2.45	0.9669	3.15
0.7	0.3	0.9755	3.43	0.9600	2.93	0.9667	3.11	0.8883	2.56	0.9669	3.27
0.8	0.7	0.9664	3.97	0.9337	3.23	0.9563	3.65	0.9281	2.99	0.9540	3.78
0.9	0.9	0.9994	4.85	0.9917	3.71	0.9967	4.56	0.9955	3.75	0.9967	4.63



a) Woolf 방법      b)  $\lambda_{ijk} = 1/8$  방법      c) (XZ, YZ) 모형      d) (XY, XZ, YZ) 모형  
<그림 4.2>  $\theta_{MH}$  가 5에서 변할 경우 각 방법에 따른 포함확률( $n_{ijk}^* = n_{ijk} + 4 \lambda_{ijk}$ )

각층의 오즈비가 다른, 일반적인 경우의 공통 오즈비에 대한 포함확률을 비교한 결과 오즈비의 변화가 없는 경우의 결과보다 포함확률이 좋지 않게 나타났다. 로그 오즈비와 표준오차가 존재하지 않는  $n_{121} = n_{212} = 0, n_{211} = n_{122} = 0, n_{i11} = n_{i22} = 0$  (단,  $i, j = 1, 2$ )인 경우와 열합계가 0인  $n_{11k} = n_{21k} = 0, n_{12k} = n_{22k} = 0$  (단,  $k = 1, 2$ )일 경우 신뢰구간의 평균길이를 계산할 수가 없으므로 이들 경우는 평균길이의 계산에서 제외하였다. <그림 4.2>에서 (XYZ)모형의 포함확률 그래프는 Woolf방법이나 (XZ, YZ)모형의 포함확률 그래프와 비슷하였다.

<표 4.2>와 <그림 4.2>에 의하면  $\lambda_{ijk}$ 의 추정량으로 (XY, XZ, YZ)모형과  $\lambda_{ijk} = 1/8$ 을 사용한 경우 포함확률은 명목수준 0.95이하로 많이 내려가며, (XZ, YZ)모형과 (XYZ)모형인 경우 포함확률은 서로 비슷하며 양호하다. 그리고 신뢰구간의 수정을 많이한 Woolf방법은 포함확률이 대부분 명목수준보다 높다. 따라서  $\lambda_{ijk}$ 으로 신뢰구간의 수정이 작고 계산이 간단한 (XZ, YZ)모형의 칸 확률값을 사용하는 것이 좋다.

#### 4.2 위험함수를 고려한 가중요소 $\hat{W}$

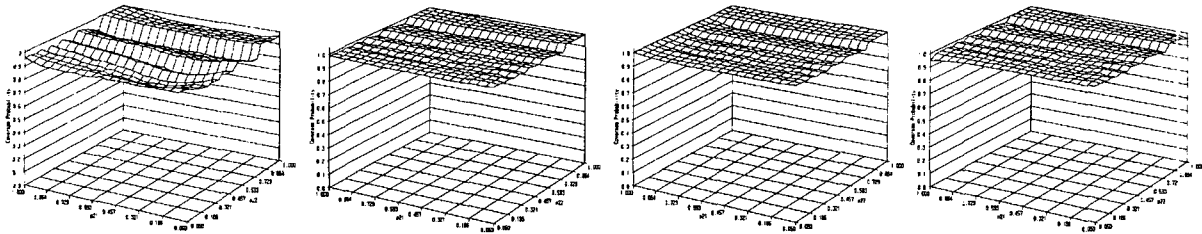
자료에 독립적인 상수 대신 자료의 정보를 이용하여 식(2.4)를 3차원으로 확장한 다음의  $\hat{W}$

$$\hat{W} = \frac{N^2 - \sum \sum \sum n_{ijk}^2}{\sum \sum \sum n_{ijk}^2 - 2N \sum \sum \sum n_{ijk} \lambda_{ijk} + N^2 \sum \sum \sum \lambda_{ijk}^2} \quad (4.8)$$

<표 4.3>  $\theta_{MH}$  가 5에서 변할 경우 로그 오즈비에 대한 포함확률과 평균길이( $n_{1k}=n_{2k}=10, k=1,2$ )

$$n_{ijk}^* = N \cdot \hat{p}_{ijk}^*$$

$p_{21}$	$p_{22}$	$\lambda_{ijk}=1/8$		(XZ, YZ)		(XY, XZ, YZ)		(XYZ)	
		포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이
0.1	0.1	0.9222	2.80	0.9784	3.63	0.9785	3.63	0.9784	3.63
0.2	0.1	0.9064	2.71	0.9760	3.35	0.9759	3.35	0.9759	3.35
0.3	0.2	0.8688	2.62	0.9700	2.97	0.9695	2.97	0.9695	2.97
0.4	0.3	0.8481	2.60	0.9662	2.92	0.9653	2.92	0.9653	2.92
0.5	0.5	0.8382	2.64	0.9575	3.10	0.9575	3.10	0.9575	3.10
0.6	0.2	0.9042	2.69	0.9765	3.30	0.9764	3.30	0.9763	3.30
0.7	0.3	0.9070	2.74	0.9755	3.43	0.9761	3.43	0.9755	3.43
0.8	0.7	0.8916	2.98	0.9664	3.97	0.9686	3.97	0.9664	3.97
0.9	0.9	0.9875	3.42	0.9994	4.85	0.9997	4.85	0.9994	4.85



a)  $\lambda_{ijk}=1/8$       b) (XZ, YZ) 모형      c) (XY, XZ, YZ)모형      d) (XYZ)모형

<그림 4.3>  $\theta_{MH}$  가 5에서 변할 경우 각 방법에 따른 포함확률( $n_{ijk}^* = N \cdot \hat{p}_{ijk}^*$ )

을 사용하며  $\lambda_{ijk}$ 의 추정량은 앞 절의 4가지 모형의 칸 확률이 사용된다.  $\hat{W}$ 는 자료에 따라 변하며  $\hat{p}$ 를  $\lambda$ 방향으로 평활화 시킨다. 유사 베イズ 추정량  $\hat{p}^*$ 를 사용한 평활화된 칸 추정값  $n_{ijk}^*$ 는

$$n_{ijk}^* = N \cdot \hat{p}_{ijk}^* \tag{4.9}$$

이다.  $2 \times 2 \times 2$  분할표에서 모수  $\theta_{MH}$ 가 변하지 않을 때  $n_{ijk}^* = N \cdot \hat{p}_{ijk}^*$ 를 사용한 결과는  $\lambda_{ijk}=1/8$ 를 사용한 경우를 제외하고는 4.1절의  $\theta_{11(k)}$ 가 고정된 경우의 결과와 거의 비슷하게 나타났으며, 두 번째 층의  $p_{12}$ 를 반올림하여 오즈비가 5에서 변하게 할 경우 결과를 <표 4.3>과 <그림 4.3>에 나타냈다.  $\hat{W}$ 는 식(4.8)을 사용하며, 로그 오즈비에 대한 신뢰구간과 평균길이가 존재하지 않는 경우는 4.1절에서와 같은 방법으로 처리했다.

두 층의 연관(association)이 달라지면  $\lambda_{ijk}=1/8$ 을 제외한 각 경우에서 칸 수정값  $N \cdot \hat{p}_{ijk}^*$ 를 이용한 방법은  $n_{ijk} + 4 \lambda_{ijk}$  방법보다 포함확률과 평균길이가 약간 커졌다. 특히 (XY, XZ, YZ)모형을 사용한 경우 포함확률이 많이 좋아졌다.  $\lambda$ 의 추정값으로 로그선형모형을 사용한 세가지 방법은 포함확률이 양호하며 서로 비슷한 결과를 제공하나 (XZ, YZ)모형인 경우 신뢰구간이 존재하지 않는 경우의 수정이 작고 계산이 간단하므로  $\lambda_{ijk}$ 에 (XZ, YZ)모형의 칸 확률을 사용한다.

4.3 유사 베이즈 추정량  $p^*$  방향으로 평활화

3.3절에서처럼 베이즈 추정량  $p_{ijk}^*$ 를 사용하여 칸도수  $n_{ijk}$ 를 다음과 같이 평활화 한다.

$$n_{ijk}^* = n_{ijk} + 4p_{ijk}^* \quad (4.10)$$

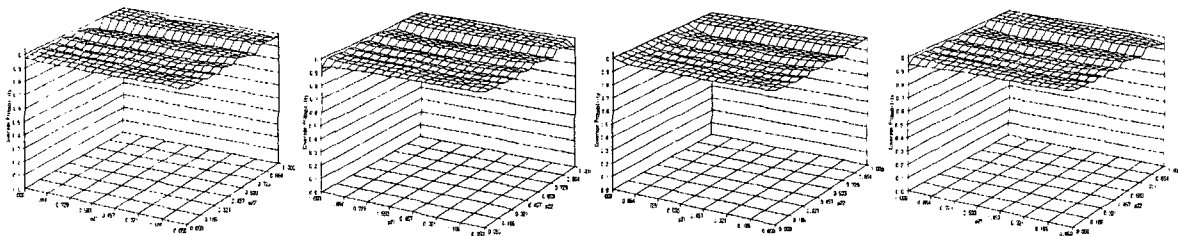
$p_{ijk}^*$ 를 구하기 위해  $\hat{W}=(1/2)t=4$  또는 식(4.8)이 사용되며  $\lambda$ 는 앞 절의 네 가지 모형의 칸 확률을 사용한다. 이 방법은 4.1절에서 평활화된  $n_{ijk}^*$ 를 구할 때  $\lambda$ 대신 베이즈 추정량  $p^*$ 를 사용하는 방법이다. 모수  $\theta_{MH}$ 가 변하지 않을 때  $\lambda_{ijk}=1/8$ 을 사용한 경우를 제외하면  $n_{ijk}^* = n_{ijk} + 4p_{ijk}^*$ 를 사용한 결과는  $n_{ijk}^* = n_{ijk} + 4\lambda_{ijk}$ 를 사용한 결과와 비슷하게 나타났다.  $p_{12}$ 를 반올림하여  $\theta_{MH}$ 가 변하게 할 경우 포함확률을 <표 4.4>와 <그림 4.4>에 나타냈다.  $2 \times 2 \times 2$  분할 표에서 유사 베이즈 추정량  $p^*$ 를 구하기 위해  $\hat{W}=(1/2)t$ 이며  $\lambda$ 은 식(4.6)을 사용했다. 로그 오즈비에 대한 신뢰구간과 평균길이 존재하지 않는 경우 4.1절에서와 같은 방법으로 처리하였다.

$\theta_{MH}$ 가 변하는 일반적인 경우에서 <표 4.4>와 4.1절의 <표 4.2>의 포함확률을 비교하면  $n_{ijk} + 4p_{ijk}^*$ 으로 평활화 시킨 결과가 더 좋으며 4.2절의  $N \cdot p_{ijk}^*$  칸수정 값을 이용한 결과보다 포함 확률이 명목수준에 더 근접한다.  $n_{ijk} + 4p_{ijk}^*$ 를 사용한 방법을 비교한 결과 신뢰구간의 수정을 최소

<표 4.4>  $\theta_{MH}$ 가 5에서 변할 경우 로그 오즈비에 대한 포함확률과 평균길이 ( $n_{1k} = n_{2k} = 10, k=1, 2$ )

$$n_{ijk}^* = n_{ijk} + 4p_{ijk}^*; \hat{W} = 4$$

$p_{21}$	$p_{22}$	$\lambda_{ijk}=1/8$		(XZ, YZ)		(XY, XZ, YZ)		(XYZ)	
		포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이	포함 확률	평균 길이
0.1	0.1	0.9719	3.40	0.9736	3.44	0.9737	3.46	0.9733	3.46
0.2	0.1	0.9688	3.15	0.9693	3.18	0.9685	3.19	0.9685	3.19
0.3	0.2	0.9568	2.82	0.9568	2.82	0.9561	2.84	0.9561	2.84
0.4	0.3	0.9523	2.77	0.9523	2.78	0.9520	2.79	0.9520	2.79
0.5	0.5	0.9444	2.93	0.9444	2.94	0.9457	2.95	0.9457	2.95
0.6	0.2	0.9680	3.11	0.9680	3.13	0.9670	3.15	0.9669	3.15
0.7	0.3	0.9682	3.23	0.9680	3.26	0.9675	3.27	0.9669	3.27
0.8	0.7	0.9523	3.71	0.9540	3.77	0.9562	3.78	0.9540	3.78
0.9	0.9	0.9949	4.50	0.9967	4.62	0.9970	4.63	0.9967	4.63



a)  $\lambda_{ijk}=1/8$  경우      b) (XZ, YZ) 모형      c) (XY, XZ, YZ) 모형      d) (XYZ) 모형  
<그림 4.4>  $\theta_{MH}$ 가 5에서 변할 경우 각 방법에 따른 포함확률 ( $n_{ijk}^* = n_{ijk} + 4p_{ijk}^*; \hat{W}=4$ )

<표 4.5>  $\theta_{MH}$ 가 5 또는 7에서 변할 경우 포함확률이 0.95미만인 비율

유사 베イズ 추정	Woolf		$\lambda_{ijk}=1/8$		(XZ, YZ)		(XY, XZ, YZ)		(XYZ)	
	$\theta_{111}=5$	$\theta_{111}=7$	$\theta_{111}=5$	$\theta_{111}=7$	$\theta_{111}=5$	$\theta_{111}=7$	$\theta_{111}=5$	$\theta_{111}=7$	$\theta_{111}=5$	$\theta_{111}=7$
$n_{ijk} + 4\lambda_{ijk}$	0.020	0.055	0.335	0.503	0.126	0.238	0.670	0.618	0.143	0.145
$N \cdot p_{ijk}^*$			0.760	0.920	0.015	0.055	0.015	0.043	0.020	0.055
$n_{ijk} + 4p_{ijk}^*$ ( $\hat{W}$ : 상수)			0.150	0.253	0.143	0.143	0.128	0.123	0.143	0.143
$n_{ijk} + 4p_{ijk}^*$ ( $\hat{W}$ : 위험함수)			0.208	0.163	0.145	0.148	0.128	0.120	0.143	0.143

로 하고 포함확률이 좋은 방법은  $\lambda$ 에 대해 조건부 독립 (XZ, YZ)모형의 칸 확률을 사용하는 것이다. 가장요인  $\hat{W}$ 에 대해  $\hat{W}=(1/2)t$ 와 위험함수를 이용하여 구한 식(4.8)을 사용했을 경우 포함확률을 비교한 결과 두 방법간에는 차이가 없었다. <표 4.5>는  $\theta_{111}=5$  또는 7이며 두 번째 층의 오즈비가 5 또는 7에서 변할 경우 4.1절~4.3절의 각 방법별 로짓 신뢰구간의 포함확률에서  $p_{21}$ 과  $p_{22}$ 에 따른 400개 포함확률 중 명목신뢰수준 0.95미만인 경우의 비율을 나타낸다. Woolf 방법은 다른 것과 비교해서 신뢰구간 수정이 많이 되었다. 따라서 네 가지 방법 중에서  $N \cdot p_{ijk}^*$ 방법이 가장 좋으며, 오즈비가 매우 크지 않을 경우 신뢰구간 수정 및 0.95미만 비율이 가장 작은 (XZ, YZ) 모형의 칸 확률사용이 추천된다.

### 5. 실제예제

<표 5.1>에 있는 Mantel(1963)의  $2 \times 2 \times 3$  자료를 이용하여 4.1절~4.3절에서 제시한 세 가지 방법을 적용한다. 이 자료는 소표본이며 0이 많이 발생하는 희박자료(sparse data)로 각 층의 연관(association)정도가 변한다. 따라서 우리는  $K=3$ 인 경우에서 유사 베イズ 추정량을 적용한 신뢰구간을 다른 방법과 비교하려 한다.  $\lambda$ 의 추정량으로 (XZ, YZ)모형에서 추정된 칸 확률을 사용하여 공동오즈비에 대한 로짓 신뢰구간을 구하여 Kim et al. (2000)에 제시된 결과와 비교하였다.

<표 5.2>에 의하면 95% 신뢰수준을 보장해 주는 정확 신뢰구간은 (2.08, 67.35)까지 신뢰구간의 폭을 좁힐 수 있으며  $N \cdot p_{ijk}^*$ 를 이용한 로짓 신뢰구간은 상한이 47.562로 더 줄었다. 이 자료에서 유사 베イズ 추정량  $p_{ijk}^*$ 를 사용하여 칸 도수를 수정한  $n_{ijk}^* = N \cdot p_{ijk}^*$ 방법이 변곡점 근사(saddlepoint approximation)방법의 결과와 거의 비슷하며 네 가지 로짓 신뢰구간 중 가장 좋다. 변곡점 근사방법은 정확검정이 가능하지 않고 일반적인 대표본 근사가 의심스러울 때 좋은 대안이다.

<표 5.1 > Penicillin 자료

Penicillin Level	Delay	Response	
		Cured	Died
1/4	None	3	3
	1 1/2 Hour	0	6
1/2	None	6	0
	1 1/2 Hour	2	4
1	None	5	1
	1 1/2 Hour	6	0



<표 5.2> 공통 오즈비에 대한 다양한 95% 신뢰구간

조건(conditional) 분포			비조건(unconditional) 분포	
변곡점근사		(1.090, 46.712)	정규근사	(1.280, 128.124)
정확(exact)	일반적(ordinary)	(1.08, 531.51)	로짓	(1.029, 34.219) <sup>1)</sup>
	수정(modified)	(2.08, 67.35)		(1.025, 47.562) <sup>2)</sup>
Mid P	일반적(ordinary)	(1.34, 266.54)		(1.124, 42.261) <sup>3)</sup>
	수정(modified)	(2.22, 56.00)		(1.132, 43.235) <sup>4)</sup>

1)  $n_{ijk} + 4 \hat{\lambda}_{ijk}$     2)  $n_{ijk}^* = N \cdot p_{ijk}^*$     3)  $n_{ijk} + 4 p_{ijk}^*$  ( $\hat{W}$ : 상수)    4)  $n_{ijk} + 4 p_{ijk}^*$  ( $\hat{W}$ : 위험함수)

## 6. 결론

본 논문에서는 소표본일 경우 유사 베イズ 추정량을 이용하여  $2 \times 2$  분할표와  $2 \times 2 \times K$  분할표에서 로그 오즈비에 대한 신뢰구간의 포함확률이 명목수준을 만족하도록 여러 가지 칸 도수 평활화 방법을 연구하였다. 칸 수정을 하지 않는 Woolf방법과 모든 칸에 동일한 수정값 0.5를 더하는 Gart방법 그리고 독립 모형으로 칸 수정을 하는 Agresti방법은 유사 베イズ 추정량을 사용한 예이다. 본 논문에서는 로짓 신뢰구간 추정에서 유사 베イズ 추정량을 이용한 칸도수 수정방법을 제안하여 비교하였다. 우리는  $2 \times 2$  그리고  $2 \times 2 \times K$  분할표에서 유사 베イズ 추정량을 이용하여 여러 가지 칸 도수를 평활화하는 방법에 따라 신뢰구간을 구축하여 로그 오즈비의 포함확률과 신뢰구간의 평균길이를 비교하였다. 우선  $2 \times 2$  분할표에서는 칸 도수  $n_{ij}$ 를 유사 베イズ 추정량  $p_{ij}^*$  방향으로 평활화 시키는  $n_{ij} + 2 p_{ij}^*$  방법이 포함확률 측면에서 가장 성능이 좋게 나타났다.  $p_{ij}^*$ 를 계산하는데  $\hat{W} = (1/2)t$ ,  $\hat{\lambda}$ 은 독립모형의 칸 확률이 추천된다.

$2 \times 2 \times K$  분할표의 경우에도 칸 수정을 하지 않는 Woolf방법과 모든 칸에 동일한 수정값 0.5를 더하는 Gart방법 이외에 유사 베イズ 추정량을 이용하기 위해 로그선형모형에 따른 조건부 독립 모형, 부분연관모형 그리고 포화모형으로 칸 도수를 평활화하는 방법으로 공통 로그 오즈비에 대한 신뢰구간의 포함확률과 평균길이를 비교하였다.  $2 \times 2 \times 2$  분할표에서는 유사 베イズ 추정량  $p_{ijk}^*$ 을 이용하여 평활화된 칸 추정값  $N \cdot p_{ijk}^*$ 으로 평활화하는 방법이 포함확률 측면에서 가장 성능이 좋았다.  $p_{ijk}^*$  계산시  $\hat{W}$ 은 위험함수를 고려한 식(4.8)이,  $\hat{\lambda}$ 은 조건부 독립 모형의 칸 확률이 추천된다. 이 결과는  $K=3$ 인 실제예제에서도 같은 결과를 얻었다. 우리는 공통오즈비 신뢰구간의 성능비교를 위해 자료 생성시 두 층의 오즈비가 같은 경우와 두 번째 층의 오즈비가 첫 번째 층의 오즈비에서 변할 경우를 각각 연구하였다. 본 논문에서의 계산은 포트란(Fortran) 77을 이용하여 작성하였다.

## 참고문헌

- [1] Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples, *Biometrics*, Vol. 55, 597-602.
- [2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

- [3] Fienberg, S. E. and Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities, *Journal of Multivariate Analysis*, Vol. 2, 127-134.
- [4] Gart, J. J. (1966). Alternative analyses of contingency tables, *Journal of the Royal Statistical Society, Series B*, Vol. 28, 164-169.
- [5] Kim, D., Hong, C. S. and Oh, M. G. (2000). The saddlepoint approximation methods for statistical inference in contingency tables, *The Korean Communications in Statistics*, Vol. 7, 313-326.
- [6] Lancaster, H. O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association*, Vol. 58, 223-234.
- [7] Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure, *Journal of the American Statistical Association*, Vol. 58, 690-700.
- [8] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, Vol. 22, 719-748.
- [9] Robins, J., Breslow, N. and Greenland, S. (1986). Estimations of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics*, Vol. 42, 311-323.
- [10] Sutherland, M., Holland, P. W. and Fienberg, S. E. (1974). Combining Bayes and frequency approaches to estimate a multinomial parameter, *In studies in Bayesian Econometrics and Statistics*, edited by S. E. Fienberg and A. Zellner, 585-617, Amsterdam, North Holland.
- [11] Woolf, B. (1955). On estimating the relation between blood group and disease, *Annals of Human Genetics*, 19, 251-253.

[ 2003년 5월 접수, 2003년 8월 채택 ]