한국통계학회논문집
2011, 18권, 4호, 445–455

DOI: 10.5351/CKSS.2011.18.4.445

오분류된 이진자료에서 Agresti-Coull 유형의 신뢰구간에 대한 이론적 고찰

이승천[1,a]

[a]한신대학교 정보통계학과

요 약

표본추출에서 오분류된 이진자료는 흔히 발생될 수 있는 현실적인 문제이지만 통계적 방법론은 상대적으로 제한적이라고 할 수 있다. 특히, 모비율의 구간추정 문제는 고전적인 Wald 방법에 의존하고 있었다. 그러나 최근 이승천과 최병수 (2009)에서 Agresti-Coull 방법을 적용하고 새로운 구간추정 방법을 제시하였으며, 수치적인 방법에 의해 Agresti-Coull 신뢰구간의 효율성을 주장하였다. 본 연구에서는 오분류된 이진자료에 대한 베이지안 모형을 다루었으며, 베이지안 모형이 Agresti-Coull 신뢰구간의 이론적 배경이 될 수 있는지 살펴 보았다.

주요용어: 오분류된 이진자료, 거짓-양성 오류, 거짓-음성 오류, 포함확률, Agresti-Coull 신뢰구간.

1. 서론

우도추정량에 의한 Wald 신뢰구간은 이항비율의 구간추정에서 하나의 표준적인 방법으로 인식되어 왔으나 Blyth와 Still (1983), Agresti와 Coull (1998), Brown 등 (2001), Lee (2006) 등 많은 연구에서 Wald 신뢰구간의 포함확률에 대한 문제가 제기되었다. 즉, Wald 신뢰구간은 대표본이론에 근거한 근사 신뢰구간으로서 표본크기가 커짐에 따라 신뢰구간의 포함확률이 점근적으로 명목 신뢰수준에 수렴하게 되어 있으나 모비율 p의 값이 0 또는 1에 가까운 경우 수렴속도가 매우 늦어 통상적인 표본크기에서는 포함확률과 명목 신뢰수준은 뚜렷한 차이를 보이게 된다. 이와 같은 Wald 신뢰구간의 특성은 비교적 오래전에 알려졌고 그동안 여러 연구에서 보다 효율적인 구간추정 방법이 대안으로 제안되었음에도 불구하고 Wald 신뢰구간은 현재까지도 비율의 구간추정을 위한 대표적인 방법으로 인식되고 있다. 이러한 이유에 대해 Brown 등 (2001)은 Wald 신뢰구간이 일반인들이 이해하기 쉬운 간편한 식에 의해 구할 수 있는 장점을 갖고 있기 때문으로 파악하고 있다. 즉, 통계적 방법의 평가에서 간편성은 효율성과 더불어 간과할 수 없는 중요한 요인이라고 할 수 있다.

Agresti-Coull 신뢰구간 (Agresti와 Coull, 1998)은 소위 "가상의 2개 성공과 2개 실패"를 관측된 자료에 추가하여 Wald 방법을 적용한 것으로 Wald 신뢰구간을 구하는 방법을 그대로 적용하기 때문에 Wald 방법과 같은 간편성을 확보하고 있을 뿐 아니라 여러 표본모형에서 매우 효율적인 구간추정 방법으로 알려져 있다. 예컨대, Brown 등 (2001)은 대안적인 구간추정 방법들을 포함확률의 근사성과 신뢰구간의 길이를 평가하여 Wilson 신뢰구간, Jeffreys 사전분포에 의한 베이지안 신뢰구간과 함께 Agresti-Coull 신뢰구간을 추천하였고, Agresti와 Caffo (2000)에서는 독립적인 이표본에서, Agresti와 Min (2005)에서는 다항분포에서 두 모비율 차이에 대한 Agresti-Coull 유형의 신뢰구간이 기존의 신뢰구간과 비교하여 포함확률의 근사성 및 신뢰구간의 길이에 있어 매우 효율적임을 보였다. 또한,

이 논문은 한신대학교 학술연구비 지원에 의하여 연구되었음.

[1] (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수. E-mail: seung@hs.ac.kr

표 1: 이중표본추출에서 각 범주의 확률

		간편 분류 0	간편 분류 1	
정밀 분류	0	n_{00} $(1-\phi)(1-p)$	n_{01} $\phi(1-p)$	$n_{0\cdot}$
	1	n_{10} θp	n_{11} $(1-\theta)p$	$n_{1\cdot}$
		$n_{\cdot 0}$	$n_{\cdot 1}$	n
		Y $1-\pi$	X π	$N-n$

Price와 Bonett (2004)에서는 독립적인 k-표본에서 Agresti-Coull 방법을 적용하여 비율들의 선형함수에 대한 효과적인 신뢰구간을 구하였다. 이 밖에 오분류된 이진자료에서도 Agresti-Coull 방법을 적용하여 효과적으로 모비율에 대한 신뢰구간을 구할 수 있었다 (Lee와 Byun, 2008; 이승천과 최병수, 2009).

모집단의 분포가 한쪽으로 치우치고 표본크기가 작은 경우, 모평균에 대한 일반적인 신뢰구간은 흔히 포함확률이 명목 신뢰수준에 못 미치게 되는 것이 일반적인데 이러한 문제를 해결하기 위해 Meeden (1999)은 가중 Polya 사전분포를 사용하여 효과적인 신뢰구간을 구축하였다. 모비율의 구간추정 문제도 모평균의 구간추정 문제와 다르지 않다. 대부분의 문제는 표본크기가 작고 p가 0 또는 1에 가까워 분포가 한쪽으로 치우친 경우에 발생하게 된다. 따라서 Meeden (1999)의 방법론은 모비율의 구간추정에도 적용될 수 있으며 이 경우 베타 사전분포를 이용한 베이지안 방법이 유용하다 (Lee, 2006).

가상적인 4개 관측값이 신뢰구간을 구하는데 왜 효과적으로 작동하는지에 대해 의문이 들 수 있겠으나 Agresti와 Coull (1998)은 베이지안 방법론을 그 이유를 설명하고 있다. 즉, 일표본 문제에서 Agresti-Coull 신뢰구간은 사실상 Beta(2, 2) 사전분포를 이용한 베이지안 신뢰구간과 같다. 즉, 1-표본, 2-표본 및 k-표본에서 Agresti-Coull 유형의 신뢰구간은 베이지안 방법론에 의해 이론적인 정당화를 할 수 있다. 본 연구에서는 이승천과 최병수 (2009)에서 제안된 Agresti-Coull 유형의 신뢰구간에 대한 이론적인 배경을 베이지안 방법론에 의해 설명하려고 한다.

2. 오분류된 이진자료에서 Agresti-Coull 신뢰구간

오분류를 갖는 이진자료는 흔히 이중표본 모형에서 발생하게 되며 여기서 고려된 이중표본 모형은 Tenenbein (1970)에서 연구된 모형과 같다. 이중표본은 두 단계의 표본추출 과정을 거치게 되는데, 첫 번째 단계는 N개의 표본을 추출하여 이를 간편 검사에 의해 양성(positive 또는 success) 또는 음성(negative 또는 failure)의 속성으로 분류한다. 이때 i-번째 단위가 양성으로 분류되면 $F_i = 1$, 음성으로 분류되면 $F_i = 0$으로 정의한다. 두 번째 단계는 N개의 표본 중 n개를 임의 추출하여 정밀검사를 하고 i-번째 단위가 양성이면 $T_i = 1$, 음성이면 $T_i = 0$으로 정의한다. 정밀검사에서는 표본이 오류없이 관찰할 수 있다고 가정한다. 따라서 모집단에서 양성의 비율은 $p = \Pr[T_i = 1]$과 같고 거짓-양성(false-positive) 오류율 ϕ와 거짓-음성(false-negative) 오류율 θ는 각각 $\phi = \Pr[F_i = 1|T_i = 0]$, $\theta = \Pr[F_i = 0|T_i = 1]$과 같이 나타낼 수 있다. 정밀 검사된 n개의 부표본(subsample)은 4개의 범주 $\{(t, f)|(0, 0), (0, 1), (1, 0), (1, 1)\}$로 구분할 수 있으며 간편 검사만으로 분류된 $N - n$개의 표본 단위는 양성과 음성으로 분류될 수 있다. 이때, 각 범주의 확률은 표 1과 같이 정리할 수 있다. 단, $\pi = \Pr[F_i = 1] = \phi(1 - p) + (1 - \theta)p$이다. 따라서 우도함수는

$$L(p, \theta, \phi) \propto [(1-\phi)(1-p)]^{n_{00}}[\phi(1-p)]^{n_{01}}[\theta p]^{n_{10}}[(1-\theta)p]^{n_{11}}\pi^x(1-\pi)^y \tag{2.1}$$

와 같다. 여기서 n_{tf}는 n개의 표본 중 범주 (t, f)에 속한 표본의 수이며, x와 y는 각각 간편 검사만으로 분류된 $N-n$개의 표본 중 양성과 음성으로 분류된 표본의 수를 나타낸다.

모형 (2.1)에서 p의 신뢰구간은 Tenenbein (1970)에 의해 구해진 p의 우도추정값

$$\hat{p} = \frac{n_{11}}{n_{\cdot 1}}\frac{x+n_{\cdot 1}}{N} + \frac{n_{10}}{n_{\cdot 0}}\frac{y+n_{\cdot 0}}{N} \tag{2.2}$$

과 우도추정량의 근사분산 추정값

$$\widehat{\mathrm{Var}}(\hat{p}) = \frac{\hat{p}^2\hat{q}^2\left(1-\hat{\phi}-\hat{\theta}\right)^2}{N\hat{\pi}\,(1-\hat{\pi})} + \frac{\hat{p}\hat{q}}{n}\left[1-\frac{\hat{p}\hat{q}\left(1-\hat{\phi}-\hat{\theta}\right)^2}{\hat{\pi}(1-\hat{\pi})}\right] \tag{2.3}$$

을 이용하여

$$\hat{p} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{p})} \tag{2.4}$$

과 같이 구할 수 있다. 단, $\hat{q} = 1-\hat{p}$, $\hat{\pi} = (x+n_{\cdot 1})/N$, $n_{\cdot f} = n_{0f} + n_{1f}$, $f = 0, 1$이고, $\hat{\phi}$과 $\hat{\theta}$는 각각

$$\hat{\phi} = \frac{n_{01}}{n_{\cdot 1}}\frac{x+n_{\cdot 1}}{N}\Big/(1-\hat{p}), \qquad \hat{\theta} = \frac{n_{10}}{n_{\cdot 0}}\frac{y+n_{\cdot 0}}{N}\Big/\hat{p}$$

과 같다.

일반적으로 식 (2.4)는 Wald 신뢰구간이라고 하는데 포함확률이 명목 신뢰수준과 비교하여 유의하게 작은 값을 갖는 등, 포함확률의 근사성에서 매우 심각한 문제가 있다 (이승천과 최병수, 2009). 또, 근사성의 문제와는 별도로 계산상의 문제도 있다. 즉, $n_{\cdot 0} = 0$ 또는 $n_{\cdot 1} = 0$이면 식 (2.2)에서 $\hat{p}$를 구할 수 없게 되는데 Tenenbein (1970)은 이 확률을 $\pi^n + (1-\pi)^n$과 같이 구하였다. 이 확률은 n이 작은 경우 결코 무시될 수 없다. 한편, $n_{0\cdot} = 0$과 $n_{1\cdot} = 0$인 경우에는 각각 $\hat{p} = 1$과 $\hat{p} = 0$이 되어 $\hat{p}$의 경우와 같이 $\hat{\phi}$ 또는 $\hat{\theta}$를 구할 수 없게 되며 n이 작을 때는 이 확률 역시 무시될 수 없다.

이승천과 최병수 (2009)는 Wald 신뢰구간이 갖고 있는 여러 가지 문제를 해결하기 위해 Agresti-Coull 유형의 신뢰구간을 구하였고, 신뢰구간의 여러 가지 특징을 모의실험을 통해 검토하였으며 이를 바탕으로 오분류를 갖는 이진자료에서 Agresti-Coull 신뢰구간이 매우 우수한 신뢰구간이라고 결론지었다. 이승천과 최병수 (2009)에서 구하여진 Agresti-Coull 신뢰구간은 가상의 4개 관측값을 표 1의 각 범주에 할당하여 $N^* = N+4$, $n^*_{tf} = n_{tf} + 0.5$, $x^* = x+1, y^* = y+1$과 같이 새로운 관측값을 구하고 이를 Wald 신뢰구간을 구하기 위한 식 (2.2), (2.3) 및 (2.4)에 대입하면 구할 수 있다.

3. 오분류된 이진자료에서 베이지안 신뢰구간

이중표본 모형에서 베이지안 방법론에 대한 많은 연구가 진행되어 왔다. 일례로 Geng과 Asano (1989)는 식 (2.1)의 자연모수(natural parameter) 대신 표 1의 각 범주확률에 대해 Dirichlet 사전분포를 적용한 베이지안 방법론을 연구하였다. 또, Raats와 Moors (2003)는 자연모수 p, θ 및 ϕ의 켤레사전분포는 베타분포이고 자연모수들이 서로 독립임을 가정하여

$$g(p, \phi, \theta) \propto p^{\alpha-1}(1-p)^{\beta-1}\phi^{\gamma-1}(1-\phi)^{\delta-1}\theta^{\epsilon-1}(1-\theta)^{\eta-1}$$

와 같은 결합사전분포를 사용하였다. 그러나 이 사전분포는 매우 복잡한 사후분포가 유도되어 앞서 언급된 간편성에서 문제가 제기될 수 있다. 즉, Raats와 Moors (2003)에서 유도된 p의 사후분포는 베타

분포들의 선형결합 형태로 나타나 p의 추론을 위해서는 매우 복잡한 계산이 요구된다. 한편, Lee와 Byun (2008)에서는 $\theta = 0$ 또는 $\phi = 0$을 가정한 이중표본 모형에서 모수변환 및 변환된 모수들의 켤레 사전분포를 이용하여 성공적으로 p의 베이지안 신뢰구간을 구축하였다.

모수변환에 의한 베이지안 방법은 상대적으로 모형을 다루기 용이하게 할 수 있으며 본 연구에서도 이 방법을 따르기로 한다. π는 $(1-p)\phi$와 $p(1-\theta)$ 두 부분으로 구성되어 있으므로 π와 $\psi = p(1-\theta)$ 또는 $(1-p)\phi$는 계층적으로 사전분포를 적용하는 것이 타당하다. 또, 주어진 π와 ψ에 대해 p는 구간 $(\psi, 1-\pi+\psi)$에서 분포하게 되므로 자연모수 대신 p, ψ, π에 대해 계층적으로 다음과 같은 사전분포를 부여한다.

$$\begin{aligned} g(p|\psi,\pi) &= \frac{1}{\mathrm{B}(\alpha,\beta)(1-\pi)^{\alpha+\beta-1}}(p-\psi)^{\alpha-1}(1-\pi+\psi-p)^{\beta-1}\mathrm{I}_{(\psi,1-\pi+\psi)}(p), \\ g(\psi|\pi) &= \frac{1}{\mathrm{B}(\gamma,\delta)\pi^{\gamma+\delta-1}}\psi^{\gamma-1}(\pi-\psi)^{\delta-1}\mathrm{I}_{(0,\pi)}(\psi), \\ g(\pi) &= \frac{1}{\mathrm{B}(\epsilon,\eta)}\pi^{\epsilon-1}(1-\pi)^{\eta-1}\mathrm{I}_{(0,1)}(\pi), \end{aligned} \tag{3.1}$$

여기서 $\mathrm{B}(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}dx$이며 $\mathrm{I}_A(x)$는 일반적인 지시함수를 나타낸다.

위에서 주어진 사전분포에 의해 p, ψ 및 π의 결합사후분포는

$$\begin{aligned} f(p,\psi,\pi|\mathcal{Y}) = &\frac{(p-\psi)^{n_{10}+\alpha-1}(1-\pi+\psi-p)^{n_{00}+\beta-1}}{\mathrm{B}(n_{10}+\alpha, n_{00}+\beta)}\frac{\psi^{n_{11}+\gamma-1}(\pi-\psi)^{n_{01}+\delta-1}}{\mathrm{B}(n_{11}+\gamma, n_{01}+\delta)} \\ &\times \frac{\pi^{x+\epsilon-\gamma-\delta}(1-\pi)^{y+\eta-\alpha-\beta}}{\mathrm{B}(x+n_{\cdot 1}+\epsilon, y+n_{\cdot 0}+\eta)}\mathrm{I}_{(\psi,1-\pi+\psi)}(p)\mathrm{I}_{(0,\pi)}(\psi)\mathrm{I}_{(0,1)}(\pi) \end{aligned} \tag{3.2}$$

와 같이 구할 수 있다. 여기서 $\mathcal{Y} = (n_{11}, n_{12}, n_{21}, n_{22}, x, y)$를 나타낸다. 이로부터 ψ와 π의 결합사후분포 및 π의 사후분포를 다음과 같이 유도할 수 있다.

$$f(\psi,\pi|\mathcal{Y}) = \frac{\psi^{n_{11}+\gamma-1}(\pi-\psi)^{n_{01}+\delta-1}}{\mathrm{B}(n_{11}+\gamma, n_{01}+\delta)}\frac{\pi^{x+\epsilon-\gamma-\delta}(1-\pi)^{y+n_{\cdot 0}+\eta-1}}{\mathrm{B}(x+n_{\cdot 1}+\epsilon, y+n_{\cdot 0}+\eta)}\mathrm{I}_{(0,\pi)}(\psi)\mathrm{I}_{(0,1)}(\pi), \tag{3.3}$$

$$f(\pi|\mathcal{Y}) = \frac{\pi^{x+n_{\cdot 1}+\epsilon-1}(1-\pi)^{y+n_{\cdot 0}+\eta-1}}{\mathrm{B}(x+n_{\cdot 1}+\epsilon, y+n_{\cdot 0}+\eta)}(\psi)\mathrm{I}_{(0,1)}(\pi) \tag{3.4}$$

즉, $\pi|\mathcal{Y} \sim \mathrm{Beta}(x+n_{\cdot 1}+\epsilon, y+n_{\cdot 0}+\eta)$이 되어 $\tilde{N} = N+\epsilon+\eta$라고 할 때 π의 사후평균 및 분산을 각각

$$\mathrm{E}(\pi|\mathcal{Y}) = \hat{\pi}_B = \frac{x+n_{\cdot 1}+\epsilon}{\tilde{N}} \tag{3.5}$$

$$\mathrm{Var}(\pi|\mathcal{Y}) = \frac{(x+n_{\cdot 1}+\epsilon)(y+n_{\cdot 0}+\eta)}{\tilde{N}^2(\tilde{N}+1)} = \frac{\hat{\pi}_B(1-\hat{\pi}_B)}{\tilde{N}+1} \tag{3.6}$$

과 같이 구할 수 있다.

p의 사후분포는 Raats와 Moors (2003)의 경우와 같이 베타분포들의 선형결합 형태로 나타나게 되지만 Raats와 Moors (2003)의 경우와는 달리 p의 사후평균 및 분산은 닫혀진 식으로 표현될 수 있다. 이를 알아 보기 위해 먼저 식 (3.3)을 식 (3.4)로 나누면

$$f(\psi|\pi,\mathcal{Y}) = \frac{\psi^{n_{11}+\gamma-1}(\pi-\psi)^{n_{01}+\delta-1}}{\mathrm{B}(n_{11}+\gamma, n_{01}+\delta)}\mathrm{I}_{(0,\pi)}(\psi)$$

과 같이 구하여져 π와 $\mathcal{Y}$ 조건부 ψ의 주변확률분포가 π에서 절삭된 베타분포임을 알 수 있으며 조건부 기대값은

$$\mathrm{E}(\psi|\pi,\mathcal{Y}) = \frac{n_{11}+\gamma}{n_{\cdot 1}+\gamma+\delta}\,\pi = \tilde{\lambda}_1\pi \tag{3.7}$$

과 같이 구하여 진다. 식 (3.7)의 $\tilde{\lambda}_1$는 $\lambda_1 = \Pr[T_1 = 1|F_i = 1] = p(1-\theta)/\pi$를 부표본으로 추정한 일종의 베이지안 추정값으로, 이 기호를 이용하면 π와 $\mathcal{Y}$ 조건부 ψ의 분산은 $\tilde{n}_{\cdot 1} = n_{\cdot 1}+\gamma+\delta$라고 할 때

$$\mathrm{Var}(\psi|\pi,\mathcal{Y}) = \frac{\tilde{\lambda}_1\left(1-\tilde{\lambda}_1\right)}{\tilde{n}_{\cdot 1}+1}\pi^2$$

과 같이 나타낼 수 있다. 그러므로 ψ의 베이지안 추정값은

$$\hat{\psi}_B = \mathrm{E}(\psi|\mathcal{Y}) = \mathrm{E}\left(\mathrm{E}(\psi|\pi,\mathcal{Y})|\mathcal{Y}\right) = \tilde{\lambda}_1\hat{\pi}_B \tag{3.8}$$

과 같다. 한편 π와 ψ의 사후공분산과 ψ의 사후분산은 각각

$$\begin{aligned}
\mathrm{Cov}(\pi,\psi|\mathcal{Y}) &= \mathrm{E}\left(\pi\mathrm{E}\left(\psi|\pi,\mathcal{Y}\right)|\mathcal{Y}\right) - \mathrm{E}\left(\pi|\mathcal{Y}\right)\mathrm{E}\left(\psi|\mathcal{Y}\right) = \tilde{\lambda}_1\left(\mathrm{E}\left(\pi^2|\mathcal{Y}\right) - \hat{\pi}_B^2\right) = \tilde{\lambda}_1\mathrm{Var}(\pi|\mathcal{Y}),\\
\mathrm{Var}(\psi|\mathcal{Y}) &= \mathrm{E}\left(\mathrm{E}\left(\psi^2|\pi,\mathcal{Y}\right)|\mathcal{Y}\right) - \tilde{\lambda}_1^2\hat{\pi}_B^2 = \mathrm{E}\left[\left(\tilde{\lambda}_1^2 + \frac{\tilde{\lambda}_1\left(1-\tilde{\lambda}_1\right)}{\tilde{n}_{\cdot 1+1}}\right)\pi^2\middle|\mathcal{Y}\right] - \tilde{\lambda}_1^2\hat{\pi}_B^2\\
&= \frac{\tilde{\lambda}_1\left(1-\tilde{\lambda}_1\right)}{\tilde{n}_{\cdot 1+1}}\left(\mathrm{Var}(\pi|\mathcal{Y}) + \hat{\pi}_B^2\right) + \tilde{\lambda}_1^2\mathrm{Var}(\pi|\mathcal{Y})
\end{aligned}$$

과 같이 구할 수 있다. 같은 방법으로 식 (3.2)와 (3.3)으로부터 p의 완전조건부 분포는

$$f(p|\psi,\pi,\mathcal{Y}) = \frac{(p-\psi)^{n_{10}+\alpha-1}(1-\pi+\psi-p)^{n_{00}+\beta-1}}{\mathrm{B}(n_{10}+\alpha,n_{00}+\beta)(1-\pi)^{n_{\cdot 0}+\alpha+\beta-1}}\mathrm{I}_{(\psi,1-\pi+\psi)}(p) \tag{3.9}$$

와 같이 구할 수 있으며 p의 완전조건부 기대값은

$$\mathrm{E}(p|\psi,\pi,\mathcal{Y}) = \frac{n_{10}+\alpha}{n_{\cdot 0}+\alpha+\beta}(1-\pi)+\psi = \tilde{\lambda}_2(1-\pi)+\psi \tag{3.10}$$

이 된다. 여기서 $\tilde{\lambda}_2$는 $\tilde{\lambda}_1$의 경우와 같이 부표본으로 구한 $\lambda_2 = \Pr[T_i = 1|F_i = 0]$의 베이지안 추정량을 나타낸다.

식 (3.5), (3.7) 및 (3.10)의 결과로부터 p의 베이지안 추정값은

$$\hat{p}_B = \mathrm{E}(p\,|\,\mathcal{Y}) = \mathrm{E}(\mathrm{E}(p|\psi,\pi,\mathcal{Y})|\mathcal{Y}) = \tilde{\lambda}_2(1-\hat{\pi}_B) + \tilde{\lambda}_1\hat{\pi}_B$$

이 됨을 알 수 있다. 한편, p의 사후분산은 잘 알려진 다음의 식을 이용하여 구할 수 있다.

$$\mathrm{Var}(p\,|\mathcal{Y}) = \mathrm{E}(\mathrm{Var}(p\,|\psi,\pi,\mathcal{Y})) + \mathrm{Var}(\mathrm{E}(p\,|\psi,\pi,\mathcal{Y})) \tag{3.11}$$

즉, $\tilde{n}_{\cdot 0} = n_{\cdot 0}+\alpha+\beta$라고 할 때,

$$\begin{aligned}
\mathrm{Var}(p\,|\psi,\pi,\mathcal{Y}) &= \mathrm{E}\left(p^2|\psi,\pi,\mathcal{Y}\right) - \{\mathrm{E}(p\,|\psi,\pi,\mathcal{Y})\}^2\\
&= \mathrm{E}\left((p-\psi)^2\,|\psi,\pi,\mathcal{Y}\right) + 2\psi\mathrm{E}\left(p|\psi,\pi,\mathcal{Y}\right) - \psi^2 - \left(\tilde{\lambda}_2(1-\pi)+\psi\right)^2\\
&= \frac{\tilde{\lambda}_2(1-\tilde{\lambda}_2)}{\tilde{n}_{\cdot 0}+1}(1-\pi)^2
\end{aligned}$$

이므로 식 (3.11)의 오른쪽 첫 번째 항은

$$\mathrm{E}(\mathrm{Var}(p\,|\psi,\pi,\mathcal{Y})) = \frac{\tilde{\lambda}_2\left(1-\tilde{\lambda}_2\right)}{\tilde{n}_{\cdot 0}+1}\left[(1-\hat{\pi}_B)^2 + \mathrm{Var}(\pi|\mathcal{Y})\right]$$

과 같이 나타낼 수 있고, 두 번째항은

$$\begin{aligned}\mathrm{Var}\left(\mathrm{E}(p\,|\psi,\pi,\mathcal{Y}\right) &= \mathrm{Var}\left(\tilde{\lambda}_2(1-\pi)+\psi|\mathcal{Y}\right)\\ &= \tilde{\lambda}_2^2\mathrm{Var}(\pi|\mathcal{Y}) - 2\tilde{\lambda}_2\mathrm{Cov}(\pi,\psi|\mathcal{Y}) + \mathrm{Var}(\psi|\mathcal{Y})\\ &= \left(\tilde{\lambda}_1-\tilde{\lambda}_2\right)^2\mathrm{Var}(\pi|\mathcal{Y}) + \frac{\tilde{\lambda}_1(1-\tilde{\lambda}_1)}{\tilde{n}_{\cdot 1}+1}\left(\mathrm{Var}(\pi|\mathcal{Y}) + \hat{\pi}_B^2\right)\end{aligned}$$

이 된다. 이 식에서 $\tilde{\lambda}_1$과 $\tilde{\lambda}_2$ 또, ϕ와 θ의 추정값은 각각 $\tilde{\lambda}_1 = \hat{\psi}_B/\hat{\pi}_B$, $\tilde{\lambda}_2 = (\hat{p}_B - \hat{\psi}_B)/(1-\hat{\pi}_B)$, $\hat{\phi}_B = (\hat{\pi}_B - \hat{\psi}_B)/(1-\hat{p}_B)$, $\hat{\theta}_B = (\hat{p}_B - \hat{\psi}_B)/\hat{p}_B$이므로 $\tilde{\lambda}_1 - \tilde{\lambda}_2$를 자연모수의 추정값으로 표현하면

$$\tilde{\lambda}_1 - \tilde{\lambda}_2 = \frac{\hat{p}_B\hat{q}_B\left(1-\hat{\phi}_B-\hat{\theta}_B\right)}{\hat{\pi}_B\left(1-\hat{\pi}_B\right)}$$

과 같다. 여기서 $\hat{q}_B = 1-\hat{p}_B$이다. 이 결과들을 정리하면 p의 베이지안 추정값과 분산을 다음과 같이 유도할 수 있다.

정리 1. 식 (3.1)에서 주어진 사전분포에 대해 p의 사후평균 및 분산은 각각 다음과 같다.

$$\hat{p}_B = \frac{n_{11}+\gamma}{\tilde{n}_{\cdot 1}}\frac{x+n_{\cdot 1}+\epsilon}{\tilde{N}} + \frac{n_{10}+\alpha}{\tilde{n}_{\cdot 0}}\frac{y+n_{\cdot 0}+\eta}{\tilde{N}}, \tag{3.12}$$

$$\begin{aligned}\mathrm{Var}(p|\mathcal{Y}) = {}& \frac{1}{\tilde{N}+1}\frac{\hat{p}_B^2\hat{q}_B^2\left(1-\hat{\phi}_B-\hat{\theta}_B\right)^2}{\hat{\pi}_B\left(1-\hat{\pi}_B\right)} + \frac{\tilde{\lambda}_1\left(1-\tilde{\lambda}_1\right)}{\tilde{n}_{\cdot 1}+1}\left(\hat{\pi}_B^2 + \frac{\hat{\pi}_B\left(1-\hat{\pi}_B\right)}{\tilde{N}+1}\right)\\ &+ \frac{\tilde{\lambda}_2\left(1-\tilde{\lambda}_2\right)}{\tilde{n}_{\cdot 0}+1}\left((1-\hat{\pi}_B)^2 + \frac{\hat{\pi}_B\left(1-\hat{\pi}_B\right)}{\tilde{N}+1}\right).\end{aligned} \tag{3.13}$$

식 (3.12), (3.13)의 결과와 베이지안 추정량의 점근적 정규성을 이용하면 다음과 같이 Wald 유형의 p에 대한 신뢰구간을 설정할 수 있다.

$$\hat{p}_B \pm z_{\alpha/2}\sqrt{\mathrm{Var}(p|\mathcal{Y})} \tag{3.14}$$

그러나 식 (3.14)를 사용하기 위해서는 사전분포의 모수값이 주어져야 한다. 사전분포의 모수는 p,ψ 및 π에 대한 사전정보를 반영하는 것이 일반적이지만 구간추정 문제에서는 흔히 무정보사전분포가 주어지며 여기에서는 균등 사전분포 ($\alpha=\beta=\gamma=\delta=\epsilon=\eta=1$), Jeffrey's 사전분포 ($\alpha=\beta=\gamma=\delta=\epsilon=\eta=1/2$), Agresti-Coull 사전분포 ($\alpha=\beta=\gamma=\delta=1/2, \epsilon=\eta=2$), 그리고 Lee (2006) 및 Lee와 Byun (2008)에서 사용된 사전분포 ($\alpha=\beta=\gamma=\delta=z_{\alpha/2}^2/8, \epsilon=\eta=z_{\alpha/2}^2/2$)에 대해 포함확률의 근사성 및 신뢰구간의 길이를 토대로 효율성을 검토하기로 한다.

위에 주어진 각 사전분포로부터 서로 다른 베이지안 신뢰구간을 구할 수 있으며 이를 각각 $CI\,[B_U]$, $CI\,[B_J]$, $CI\,[B_A]$, $CI\,[B_L]$으로 표시하고 Agresti-Coull 유형의 신뢰구간은 $CI\,[A]$으로 나타내기로 한다.

4. Agresti-Coull 유형의 신뢰구간과 베이지안 신뢰구간의 관계

본 연구의 동기는 Agresti-Coull 유형의 신뢰구간에 대한 이론적 배경을 베이지안 방법에 의해 설명하는데 있다. 즉, Agresti-Coull 유형의 신뢰구간이 사실상 베이지안 신뢰구간이라는 것을 규명하는 것이 주 목적이다. 이를 위해 두 유형의 신뢰구간에 대해 보다 자세히 살펴보기로 한다.

먼저 식 (2.2)와 (3.12)는 매우 비슷한 식으로 베이지안 신뢰구간을 위한 사전분포가 Agresti-Coull 사전분포라고 할 때, 자연모수들의 베이지안 추정값들은 2절에서 기술된 Agresti-Coull 유형의 추정값들과 일치하게 된다. 한편, 분산의 추정식에서도 식 (2.3)과 (3.13)의 첫 번째 항은 매우 유사하다. 그러나 나머지 항에서 식의 차이를 보이고 있다.

Tenenbein (1970)은 $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\pi}$를 해당 모수의 우도추정량이라고 할 때, 델타(delta) 방법을 적용하여 $\hat{p}$의 근사 분산을

$$\begin{aligned}\mathrm{Var}(\hat{p}) &\doteq \pi^2 \mathrm{Var}\left(\hat{\lambda}_1\right) + (1-\pi)^2 \mathrm{Var}\left(\hat{\lambda}_2\right) + (\lambda_1 - \lambda_2)^2 \mathrm{Var}(\hat{\pi}) \\ &= \frac{1}{n\pi}\lambda_1(1-\lambda_1)\pi^2 + \frac{1}{n(1-\pi)}\lambda_2(1-\lambda_2)(1-\pi)^2 + \frac{p^2q^2(1-\phi-\theta)^2}{N\pi(1-\pi)} \qquad (4.1)\end{aligned}$$

과 같이 전개하고 식에 나타난 모수의 값을 우도추정값으로 대치하여 식 (2.3)을 구하였다. 여기서 마지막 항인 $(\lambda_1 - \lambda_2)^2\mathrm{Var}(\hat{\pi})$의 추정값은 식 (2.3)과 (3.13)의 첫 번째 항에 유사한 형태로 나타난다. 식 (4.1)의 나머지 두 개의 항을 우도추정할 경우 분산식은 (2.3)과 같이 구할 수 있는 반면, 이를

$$\mathrm{E}\left(\pi^2|\mathcal{Y}\right)\mathrm{Var}(\lambda_1|\mathcal{Y}) + \mathrm{E}\left((1-\pi)^2|\mathcal{Y}\right)\mathrm{Var}(\lambda_2|\mathcal{Y})$$

와 같이 베이지안 추정을 하게 되면 식 (3.13)과 같이 분산식을 구할 수 있다. 또, 식 (2.1)의 모형에서 $\mathrm{E}(n_{\cdot 1}) = n\pi$, $\mathrm{E}(n_{\cdot 0}) = n(1-\pi)$이므로 식 (4.1)의 $n\pi$와 $n(1-\pi)$를 우도추정하지 않고 $n_{\cdot 1}$과 $n_{\cdot 0}$으로 적률추정을 하게 되면

$$\frac{\hat{\lambda}_1\left(1-\hat{\lambda}_1\right)}{n_{\cdot 1}}\hat{\pi}^2 + \frac{\hat{\lambda}_2\left(1-\hat{\lambda}_2\right)}{n_{\cdot 0}}(1-\hat{\pi})^2$$

와 같이 구할 수 있어 $O(N^{-1})$항을 무시하면 식 (3.13)의 마지막 두 항과 유사한 식을 갖게 된다. 결과적으로 Agresti-Coull 신뢰구간을 구하기 위한 분산추정값은 N이 클 경우, 베이지안 방식에 의한 분산식과 매우 유사하다는 것을 알 수 있다.

예제 1. Raats와 Moors (2003)에 의하면 네델란드에서는 일 년에 약 100억 유로가 사회보장 비용으로 지출이 되고 있으며 이를 6개의 회사에서 관장하고 있다고 한다. 그런데 네델란드의 사회보장제도는 규정이 매우 복잡한 것으로 유명해, 이 분야의 전문가라고 할지라고 규정을 잘못 적용하기 쉽기 때문에 일 년에 약 1억 5000만 유로 정도가 잘못 집행이 되고 있는 것으로 추측하고 있다. 이를 시정하기 위해 한 회사의 회계감사관이 자사에서 집행된 건수 중 500건을 임의로 추출하여 조사한 결과 16건에서 오류가 있다는 것을 발견하였다. 또, 감독기관에서는 이 결과를 재확인하기 위해 500건 중 53건을 표본으로 추출하여 정밀 분석을 한 결과 회계감사관의 검사에서도 오류가 있음을 발견하였다. 회계감사관과 감독기관에서 조사한 결과를 요약하면 $n_{00} = 50, n_{01} = 1, n_{10} = 0, n_{11} = 2, x = 14, y = 433$과 같다.

상기의 관찰값과 4개의 가상 관찰값을 더하여 계산된 p의 추정값과 추정분산은 각각 $\hat{p}^* = 0.0689$, $\widehat{\mathrm{Var}}(\hat{p}^*) = 9.3714E{-}4$. 또, 각 사전분포에 대해 $(\hat{p}_B, \mathrm{Var}(p|\mathcal{Y}))$를 계산하면 $CI[B_U], CI[B_J], CI[B_A], CI[B_L]$에 대해 각각 $(0.0750, 9.6109E-4), (0.0673, 8.5757E-4), (0.0689, 8.6427E-4), (0.0685, 8.5978E-4)$와 같이 구할 수 있었다. 이 결과를 이용하여 p의 95% 신뢰구간을 구한 결과가 표 2에 나와 있다.

표 2: Raats와 Moors (2003)의 사회보장제도 자료에서 p의 95% 신뢰구간

신뢰구간의 종류	하한	상한	구간의 길이
$CI[A]$	0.0089	0.1289	0.1200
$CI[B_U]$	0.0142	0.1358	0.1216
$CI[B_J]$	0.0099	0.1247	0.1148
$CI[B_A]$	0.0112	0.1265	0.1153
$CI[B_L]$	0.0110	0.1260	0.1150

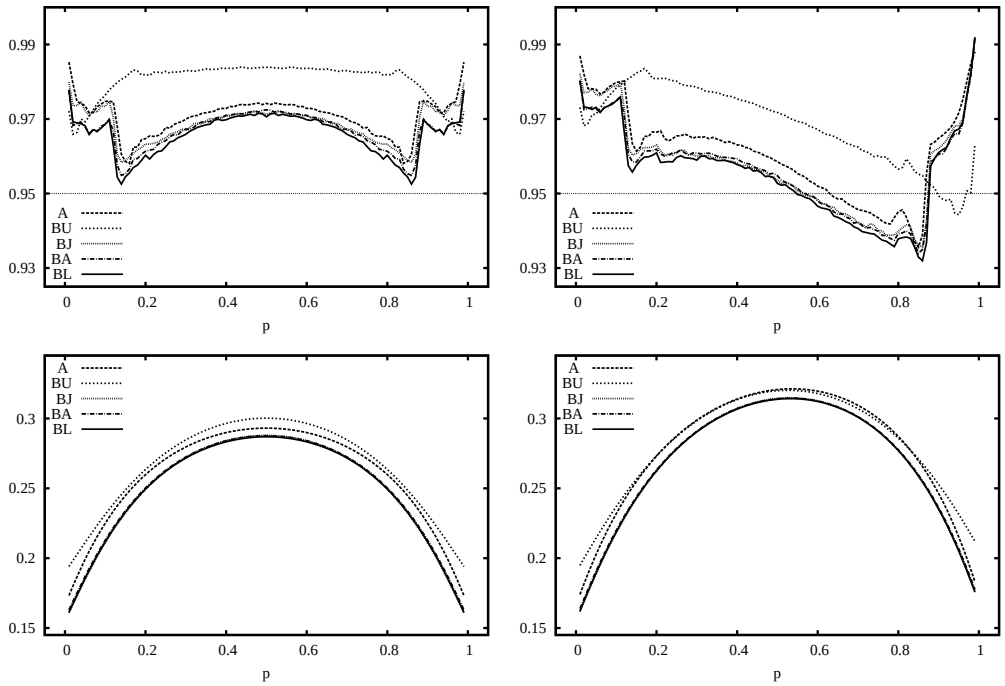

그림 1: *95% 신뢰구간들의 포함확률(상단)과 기대길이(하단),* $(\phi,\theta) = (0.1,0.1)$*(왼쪽),* $(0.1,0.2)$*(오른쪽),* $N = 200, n = 20$*일 때*

$CI[A]$의 길이는 다른 베이지안 신뢰구간의 길이와 큰 차이를 보이지 않고 있으나 $CI[B_U]$를 제외하면 가장 신뢰구간의 폭이 넓은 것을 관찰할 수 있다. 특히, $CI[A]$는 $CI[B_A]$와 같은 p의 추정값을 갖고 있으나 구간의 폭이 약간 넓은 것을 관찰할 수 있다. 일반적으로 Agresti-Coull 유형의 신뢰구간은 보수적인 경향을 띄게 되는데 베이지안 방법은 Agresti-Coull 신뢰구간의 보수성을 일부 수정한 것으로 판단된다.

5. 신뢰구간의 비교

이 절에서는 앞 절에서 언급된 5개의 신뢰구간에 대해 포함확률의 근사성 및 구간의 길이를 비교하기로 한다. 대표본 이론에 의하면 우도추정량과 베이지안 추정량은 점근적으로 정규성을 갖게 되므로 각 신뢰구간의 포함확률은 점근적으로 명목 신뢰수준으로 수렴하게 되어 있다. 그러므로 여기서 표본크기가 작은 경우의 효율성을 비교하기로 한다.

비교를 위해 $N = 200$, $n = 0.1N\{(\phi,\theta)| = (0.1,0.1),(0.1,0.2)\}$이고, p는 0.01의 간격으로 $(0.01,0.99)$

구간의 모든 값에서 포함확률과 기대길이를 계산하였다. 거짓-양성 및 거짓-음성 오류는 그 값이 크지 않는 것이 일반적이기 때문에 ϕ와 θ의 값은 0.1 또는 0.2인 경우만으로 고려하여 그림 1의 결과를 얻게 되었다. 만약 $T_i = 0$를 성공 사건이 발생한 것으로 생각하면 거짓-양성 오류와 거짓-음성 오류는 그 역할이 바뀌게 된다. 그러므로 그림 1에서 $(\phi, \theta) = (0.2, 0.1)$인 경우는 오른쪽 그림의 p를 q로 해석하여 추론될 수 있다.

그림에서 $CI[B_U]$는 대부분의 p 값에서 가장 넓은 기대길이를 갖으며 결과적으로 매우 보수적인 경향을 띄고 있다. 그러나 p가 0 또는 1에 가까운 경우, 가장 넓은 구간의 길이에도 불구하고 포함확률은 다른 구간과 비교하여 작은 값을 갖는 등, 다른 4개의 신뢰구간과 비교하여 다른 경향을 보이고 있다. 이를 근거로 결론을 내린다면 균등 사전분포는 p의 구간추정 문제에는 적절치 않은 것으로 판단된다.

$CI[B_U]$를 제외한 나머지 4개의 신뢰구간은 포함확률 및 기대길이에 있어 매우 유사한 패턴을 보이고 있다. 이는 $1 - \alpha = 0.95$일 경우, $z^2_{\alpha/2} = 1.96^2 \approx 4$이므로 $CI[B_L]$와 $CI[B_A]$ 사실상 같다고 볼 수 있으며, ϵ과 η는 전체 표본에 추가된 가상적인 성공과 실패의 수로 해석될 수 있는 반면, 다른 사전분포의 모수들은 부표본에 추가된 가상 관찰값의 수로 해결이 될 수 있다. 그러므로 N이 n과 비교하여 상대적으로 클 때, ϵ과 η의 역할은 그다지 크지 않다. 즉, $CI[B_J], CI[B_A], CI[B_L]$ 및 $CI[A]$는 사실상 큰 차이가 없다고 하겠다.

6. 결론

이중표본 추출은 표본추출 비용을 줄일 수 있다는 점에서 매우 실용적인 표본추출 방법으로 국가단위의 대규모 표본조사나 의학실험과 같이 고가의 표본조사 비용이 요구되는 경우 유용하게 사용될 수 있다. 그러나 이중표본 추출에서는 대부분의 조사단위과 거짓-양성과 거짓-음성 오류에 노출되어 있어 통계적 분석은 상대적으로 어렵다고 할 수 있다. 특히, 모수의 구간추정은 통계적 추론을 위해 핵심적인 역할을 담당하고 있으나 이중표본 추출에서와 같이 오분류된 이진자료로 비율의 구간추정을 다룬 연구 결과는 많지 않았으며, 지금까지 Tenenbein (1970)에 의해 유도된 우도추정량을 이용한 Wald 신뢰구간이 주로 사용되고 있다. 그러나 Wald 신뢰구간은 여러 표본모형에서 오류가 많은 구간추정 방법으로 정밀한 통계적 분석이 요구될 경우에서 사용될 수 없다. 이와 비교하여 Agresti-Coull 신뢰구간은 간편성 및 효율성에서 우수한 통계적 방법으로 여러 가지 표본모형에 쉽게 적용될 수 있다는 장점을 있다. 특히, 오분류된 이진자료에서도 포함확률의 근사성 및 구간의 길이 측면에서 매우 효과적인 통계적 방법임이 밝혀 졌다 (이승천과 최병수, 2009). 그러나 Agresti-Coull 신뢰구간이 왜 효율적인지에 대해서는 밝혀진 바가 없다.

Agresti와 Coull (1998)에서는 일표본 문제에서 Agresti-Coull 신뢰구간의 이론적 배경을 베이지안 방법론을 설명하고 있으나 오분류된 이진자료에서도 적용이 가능한지에 대해서는 알려진 바가 없기 때문에 본 연구에서는 오분류된 이진자료의 베이지안 모형을 설정하고 이 모형에 의해 Agresti-Coull 신뢰구간이 설명될 수 있는지 살펴 보았다. 결론적으로 Agresti-Coull 신뢰구간은 무정보 사전분포에 근거한 베이지안 신뢰구간과 사실상 차이가 없었으며, 간편성에 있어서는 오히려 베이지안 신뢰구간보다 우수하다고 할 수 있었다.

참고 문헌

이승천, 최병수 (2009). 이중표본에서 모비율의 구간추정, <응용통계연구>, **22**, 1289–1300.

Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures, *The American Statistician*, **54**, 280–288.

Agresti, A. and Coull, B. A. (1998). Approximation is better than "exact" for interval estimation of binomial proportions, *The American Statistician*, **52**, 119–126.

Agresti, A. and Min, Y. (2005). Simple improved confidence intervals for comparing matched proportions, *Statistics in Medicine*, **24**, 729–740.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals, *Journal of the American Statistical Association*, **78**, 108–116.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion, *Statistical Science*, **16**, 101–133.

Geng, Z. and Asano, C. (1989). Bayesian estimation methods for categorical data with misclassifications, *Communications in Statistics, Theory and Methods*, **18**, 2935–2954.

Lee, S.-C. (2006). Interval estimation of binomial proportions based on weighted Polya posterior, *Computational Statistics & Data Analysis*, **51**, 1012–1021.

Lee, S.-C. and Byun, J.-S. (2008). A Bayesian approach to obtain confidence intervals for binomial proportion in a double sampling scheme subject to false-positive misclassification, *Journal of the Korean Statistical Society*, **37**, 393–403.

Meeden, G. D. (1999). Interval estimators for the population mean for skewed distributions with a small sample size, *Journal of Applied Statistics*, **26**, 81–96.

Price, R. M. and Bonett, D. G. (2004). An improved confidence interval for a linear function of binomial proportions, *Computational Statistics & Data Analysis*, **45**, 449–456.

Raats, V. M. and Moors, J. J. A. (2003). Double-checking auditors: A Bayesian approach, *The Statistician*, **52**, 351–365.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications, *Journal of the American Statistical Association*, **65**, 1350–1361.

2011년 4월 접수; 2011년 5월 채택

Theoretical Considerations for the Agresti-Coull Type Confidence Interval in Misclassified Binary Data

Seung-Chun Lee[1,a]

[a]Department of Statistics, Hanshin University

Abstract

Although misclassified binary data occur frequently in practice, the statistical methodology available for the data is rather limited. In particular, the interval estimation of population proportion has relied on the classical Wald method. Recently, Lee and Choi (2009) developed a new confidence interval by applying the Agresti-Coull's approach and showed the efficiency of their proposed confidence interval numerically, but a theoretical justification has not been explored yet. Therefore, a Bayesian model for the misclassified binary data is developed to consider the Agresti-Coull confidence interval from a theoretical point of view. It is shown that the Agresti-Coull confidence interval is essentially a Bayesian confidence interval.

Keywords: Misclassified binary data, false-positive error, false-negative error, coverage probability, Agresti-Coull confidence interval.

This work was supported by Hanshin University research grant.

[1] Professor, Department of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do 449-791, Korea.
E-mail: seung@hs.ac.kr