응용통계연구 (2009)
22(6), 1289–1300

이중표본에서 모비율의 구간추정

이승천[1] · 최병수[2]

[1]한신대학교 정보통계학과, [2]한성대학교 멀티미디어학과

(2009년 8월 접수, 2009년 10월 채택)

요 약

표본추출 비용의 절감을 위해 흔히 사용되는 이중표본추출방법은 대부분의 표본들이 2종류의 오류에 의해 오염이 되어 있어 통계적 분석이 상대적으로 용이하지 않다. 특히, 비율의 추론을 위한 중요한 분석 도구인 구간추정은 현재까지 우도추정량의 정규근사에 의존하는 Wald 방법만이 알려져 있으나 Wald 신뢰구간은 포함확률의 근사성 등에서 많은 문제가 있다는 것이 여러 연구에서 확인되고 있다. 본 연구에서는 이중표본추출에서 Wald 신뢰구간의 문제점을 파악하고 이에 대한 대안으로 Agresti-Coull 유형의 신뢰구간을 제시한다.

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

1. 서론

이중표본추출 방법은 표본단위의 특성을 조사할 때 정밀한 검사방법을 사용하면 많은 비용이 요구되거나, 대규모의 정밀한 검사가 불가능할 경우 흔히 사용되는 표본추출방법이다. 예컨대 York 등 (1995)은 노르웨이에서 다운증후군(Down's syndrome)을 앓고 있는 신생아의 비율을 추정하기 위해 1985–1988년에 태어난 모든 신생아에 대해 외견을 조사(visual inspection)하여 다운증후군의 여부를 검사하였다. 그러나 외견에 의한 검사는 오류가 있을 수 있으므로 일부의 신생아에 대해서는 유전자 검사에 의한 정밀한 검사를 실시하는 이중표본추출방법을 사용한 바가 있고, Raats와 Moors (2003)도 네덜란드에서 사회보장(연금, 의료보험, 실업수당 등)이 필요한 비율을 조사하기 위해 이중표본추출방법을 활용한 바 있다. 즉, Raats와 Moors에 의하면 네덜란드에서는 1년에 100억 유로 이상이 사회보장 비용으로 지불되고 있고, 이를 6개의 회사에서 모두 관장한다고 한다. 그러나 네덜란드에서 사회보장과 관련된 법률과 규정은 매우 복잡한 것으로 유명하여 6개 회사에서 다루는 대규모의 사회보장 집행은 오류를 포함할 가능성이 매우 높다. 그러므로 일부의 표본에 대해 보다 정밀한 검증하여 이중표본추출 방법을 사용하였다.

이와 같이 이중표본추출방법은 오류가 있을 수 있지만 모든 표본에 대해 적용이 가능하거나 또는 적은 비용이 소요되는 검사를 적용하고, 일부의 표본에 대해서만 상대적으로 많은 비용이 요구되나 오류가 없는 정밀한 검사를 실시하는 현실적인 방법으로 대규모의 조사에서나 정밀한 검사에 많은 비용이 소요되는 것이 일반적인 의학실험 등에서는 매우 유용한 표본추출방법이라고 하겠다.

이 논문은 2008년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2008-313-C00131).
제 2저자는 2008년도 한성대학교 교내연구비에 의해 지원을 받았음.
[1]교신저자: (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수. E-mail: seung@hs.ac.kr

이중표본추출방법에서 대부분의 표본은 부정확한 검사에 의해 분류되어 2종류의 오류에 노출되어 있다. York 등 (1995)의 예를 들면, 정상적인 신생아를 다운증후군으로 분류(거짓-양성)하거나 다운증후군인 신생아를 정상아로 분류(거짓-음성)될 수 있다. 이와 같이 오류에 노출된 표본을 모수의 추론에 사용할 경우 보통의 추정방법은 매우 편의(bias)된다는 것이 잘 알려져 있다 (Bross, 1954).

한편 정밀한 검사에 의한 표본의 수는 많지 않은 경우가 대부분이며 적은 수의 표본에 의한 비율의 통계적 추론은 정밀도에서 많은 문제가 있다는 것도 잘 알려져 있다. 즉, 많은 표본조사에서 자료 분석의 중요한 도구의 하나는 비율의 신뢰구간인데 특히 의학실험에서는 비율 또는 비율의 차이에 대한 추론을 염두에 두고 조사가 이루어지는 경우가 많다. 이때 비율의 추론을 위해 흔히 Wald 신뢰구간을 이용하고 있다. Wald 신뢰구간은 중심극한정리에 기초한 신뢰구간으로 통계학 교재에는 Wald 신뢰구간을 사용할 수 있는 조건으로

1. $np, n(1-p) \geq 5$ (또는 10)
2. $np(1-p) \geq 5$ (또는 10)
3. $n\hat{p}, n(1-\hat{p}) \geq 5$ (또는 10)
4. $\hat{p} \pm 3\sqrt{\hat{p}(1-\hat{p})}$의 구간에 0 또는 1이 포함되지 않는다.
5. $n \geq 50$.

등과 같은 조건들을 나열하고 있다. 이러한 조건들은 표본크기 n이 어느 정도 이상이 되어야 한다는 것을 의미한다. 그러나 최근 Agresti와 Coull (1998), Agresti와 Caffo (2000), Brown 등 (2001), Lee (2006) 등 여러 학자들에 의해 비율 또는 비율 차이의 신뢰구간 추정에 많이 사용되고 있는 Wald 신뢰구간의 문제점이 재조명되고 있다. 즉, 최근의 연구 결과에 의하면 표본크기가 충분히 크다고 하여도 Wald 신뢰구간은 포함확률(coverage probability) 및 구간의 길이에 있어 만족할 만한 결과를 제시하지 못한다고 한다. 그러므로 이중표본추출방법에 의한 데이터를 충분히 활용되기 위해서는 두 종류의 표본을 모두 사용할 수 있는 추론 방법이 심도 깊게 연구되어야 한다.

2. 이중표본추출 모형

2.1. 일반적인 이중표본추출 모형

이중표본추출방법의 1 단계에서는 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. 2 단계에서는 N개의 표본 중 n개를 임의 추출하여 정밀 검사를 하고, i-번째 단위가 양성이면 $T_i = 1$, 음성이면 $T_i = 0$으로 정의한다. 정밀검사에서 표본을 오류없이 관찰할 수 있다고 가정하면 모집단에서 양성의 비율은 $p = \Pr[T_i = 1]$과 같고, 거짓-양성(false-positive) 오류율 ϕ와 거짓-음성(false-negative) 오류율 θ는 각각

$$\phi = \Pr[F_i = 1 | T_i = 0]$$
$$\theta = \Pr[F_i = 0 | T_i = 1]$$

와 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 2.1과 같이 정리된다. 단, $\pi = \Pr[F_i = 1] = \phi(1-p) + (1-\theta)p$이다. 따라서 우도함수는

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

표 **2.1.** 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	$N-n$

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 Tenenbein (1970)은 다음과 같이 우도추정량과

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

$$\hat{\theta} = \frac{n_{10}}{n_{\cdot 0}} \frac{Y + n_{\cdot 0}}{N} \Big/ \hat{p} \tag{2.3}$$

$$\hat{\phi} = \frac{n_{01}}{n_{\cdot 1}} \frac{X + n_{\cdot 1}}{N} \Big/ (1 - \hat{p}) \tag{2.4}$$

모비율 추정량 $\hat{p}$의 근사 분산을 구하였다.

$$\mathrm{Var}(\hat{p}) = \frac{pq}{n}\left[1 - \frac{pq(1-\theta-\phi)^2}{\pi(1-\pi)}\right] + \frac{p^2q^2(1-\theta-\phi)^2}{N\pi(1-\pi)}. \tag{2.5}$$

일반적으로 주 관심대상인 모비율 p의 추론은 식 (2.2)에 의존한다. 우도추정량의 효율성은 이미 잘 알려져 있기 때문에 Tenenbein 이후 점추정(point estimation)에 대한 연구는 더 이상의 진전이 없는 것으로 알고 있다. 따라서 베이지안 추정 방법 등과 비교하여 효율성에 대한 연구가 필요하다고 판단된다. 그러나 보다 문제가 되고 있는 부분은 구간추정이다. 즉, 현재까지 모형 (2.1)에서 모비율 p의 구간추정은 우도추정량과 우도추정량의 표준오차에 근거한 Wald 신뢰구간만이 알려져 있는 형편이다. 그러나 앞에서 언급된 바와 같이 Wald 신뢰구간은 매우 오류가 많은 구간추정 방법으로 여러 연구에서 Wald 신뢰구간의 문제점과 새로운 구간추정방법이 연구되고 있다. 예컨대, 이항분포(1-표본)에서 모비율의 구간추정에 대해 Wald 신뢰구간의 문제점을 개선한 새로운 신뢰구간이 제시되었고 (Agresti와 Coull, 1998; Lee, 2006), 독립적인 두 이항분포의 모비율 차이(2-표본)에 대한 Wald 신뢰구간의 오류를 개선한 신뢰구간이 Agresti와 Caffo (2000), 이승천 (2006)에서 제시되었다. 또, 보다 일반적인 k-표본에서는 Price와 Bonett (2004), 이승천 (2007) 등의 연구를 참조할 수 있다. 이러한 연구를 참조하면 2종의 오류에 의해 오염된 데이터를 갖게 되는 이중표본에서 Wald 신뢰구간은 구간추정량으로서의 효율성이 매우 좋지 않을 것을 쉽게 예상할 수 있다.

2.2. 거짓-양성 오류 이중표본추출 모형

이중표본추출 모형에서 보다 많은 연구가 이루어진 것은 거짓-양성 오류만을 가정한 모형이다. 즉, $\theta =$

0을 가정한 모형으로 Swaen 등 (2001), Braunstein (2002), Kazemi 등 (2001), Boese 등 (2006) 및 Lee (2007)의 연구에서 찾아 볼 수 있다. 이 모형에서는 조사항목의 자연적 특성이나 실험방법에 의해 간편검사에서의 거짓-음성 오류가 발생되지 않거나 최소화되었다는 것을 가정한다. 이 경우 $n_{10} = 0$이 되며, 모형 (2.1)은

$$L(p, \phi) \propto [(1-\phi)(1-p)]^{n_{00}} [\phi(1-p)]^{n_{01}} p^{n_{11}} \pi^{x} (1-\pi)^{y} \tag{2.6}$$

와 같이 나타낼 수 있다.

(2.6)의 모형에서 Boese 등 (2006)은 로그 프로파일 가능도(log profile-likelihood), Rao 스코어 (score) 등 우도함수와 관련된 통계량을 이용하여 모비율에 대한 5개의 구간추정 방법을 제시하였다. 이 연구에서는 5개의 구간추정량이 빈도학파적 입장에서 제시된 최초의 신뢰구간들이라는 것을 암시적으로 언급하였으며, Monte Carlo 방법에 의해 5개 신뢰구간과 Wald 신뢰구간의 포함확률을 비교 조사하여 이를 근거로 Rao score에 의한 신뢰구간을 추천하였다. 이 연구에서도 Wald 신뢰구간의 포함확률은 명목 신뢰수준(nominal confidence level)과 현격한 차이를 보이고 있어 Wald 신뢰구간을 이용한 통계적 추론은 많은 문제가 있음을 보이고 있다.

Boese 등 (2006)에 의해 추천된 $(1-\alpha) \times 100\%$ 신뢰구간은

$$S_n[I_{pp|\phi}] = \left[U_p(p, \hat{\phi}_p)\right]^2 I^{-1}_{pp|\phi} \leq \chi^2_{1,\alpha} \tag{2.7}$$

을 만족하는 p의 집합이다. 단,

$$\hat{\phi}_p = [2(N-n_{11})(1-p)]^{-1} \Big[(X+n_{01})(1-p) - (N-X-n_{11})p + \big\{ [p(X+Y+n_{00}) - (X+n_{01})]^2 + 4n_{01}p(n_{00}+Y) \big\}^{\frac{1}{2}} \Big],$$

$$I_{pp|\phi} = I_{pp} - I^2_{p\phi} I^{-1}_{\phi\phi}, \qquad I_{pp} = \frac{n}{(1-p)} + \frac{n}{p} + \frac{(N-n)(1-\hat{\phi}_p)^2}{(1-p)\hat{\phi}_p + p} + \frac{(N-n)(1-\hat{\phi}_p)}{1-p},$$

$$I_{p\phi} = \frac{N-n}{(1-p)\hat{\phi}_p + p}, \qquad I_{\phi\phi} = (1-p)\left[\frac{N}{1-\hat{\phi}_p} + \frac{n}{\hat{\phi}_p} + \frac{(N-n)(1-p)}{p + (1-p)\hat{\phi}_p} \right]$$

과 같이 주어진다. Boese 등 (2006)의 식에서는 I_{pp}와 $I_{\phi\phi}$의 식에 오류가 있어 $S_n[I_{pp|\phi}]$의 평가가 적절히 되었는가에 대한 의문이 있다. 그러나 신뢰구간의 효율성 문제와는 별도로 Boese 등 (2006)에 의해 제시된 신뢰구간들은 기본적으로 대표본에서 검정에 기반을 둔 신뢰구간으로 구간의 상한과 하한을 구하기 위한 식을 제시할 수 없다는 문제가 있다. 즉, 그들의 표현에 의하면 컴퓨터에 의해 매우 큰 행렬 연산을 수행하거나 또는 반복 실행(trial and error) 방법에 의존한다. 이와 같은 계산상의 특성은 Agresti와 Coull (1998) 또는 Brown 등 (2001)에서 강조된 구간추정에서 신뢰구간의 간편성과는 매우 동떨어져 있어 실용성에서 문제가 제기될 수 있으며 더욱 문제가 되는 것은 Lee (2007)에서 살펴볼 수 있듯이 표본크기가 상당히 큰 경우라고 하더라고 신뢰구간의 실제 포함확률이 명목신뢰수준과는 유의한 차이를 보이고 있다는 것이다.

간편성의 문제는 실제 구간을 구하기 어려운 것은 물론, 구간추정량의 특성을 살펴보는데도 많은 장애가 된다. Boese 등 (2006)의 연구에서도 구간의 포함확률에 대해 조사하였을 뿐, 구간의 길이 등 다른 중요한 특성치에 대한 언급이 없어 구간추정량으로서의 효율성을 판단하기에 미흡한 점이 있다. 실제로 Lee (2007)의 연구에 의하면 식 (2.7)의 신뢰구간은 p가 작은 경우(대략 $p \leq 0.2$) 포함확률이 매우 안정적인 것으로 나타났으나 p가 커짐에 따라 신뢰구간의 길이가 필요 이상으로 커져 매우 보수적인 신뢰구간이 되고 있다. 또 p가 1에 가까워지면 포함확률이 급격히 떨어지는 매우 불안정한 모습을 보이고

있다는 것을 발견하였다. 즉, 실제의 p 값이 크면 추천된 신뢰구간은 성능 면에서도 효율적이지 못하였다. 이러한 결과는 장애모수인 ϕ 때문인 것으로 판단된다. 즉, ϕ의 존재 때문에 p의 구간추정에서 식 (2.7)과 같은 매우 복잡한 구간추정량이 유도되었을 뿐 아니라 추정량의 효율을 떨어뜨리는 요인으로 작용하고 있다.

모형 (2.6)에서 p의 구간추정에 대한 또 다른 흥미로운 연구결과는 Lee (2007)에서 볼 수 있다. 이 연구에 의하면 모형 (2.6)에 대해 Agresti와 Coull (1998) 방법을 적용하여 보다 효율적인 신뢰구간을 구할 수 있었다. Agresti와 Coull 방법은 Wald 신뢰구간을 가상적인 4개 자료 (2개의 성공과 2개의 실패)를 추가하여 간단하게 수정한 것으로 Agresti와 Coull (1998), Agresti와 Caffo (2000), Agresti와 Min (2005)에서 제시된 바와 같이 비율, 비율 차이 및 비율의 선형결합에 대한 신뢰구간에서 매우 효과적인 신뢰구간을 제시하여 주고 있다. 이중표본추출에 있어서도 Tenenbein (1970, 1971, 1972)에 의해 제시된 우도 추정량과 Wald 방법을 이용하여 Agresti-Coull 유형의 신뢰구간은 어렵지 않게 유도할 수 있었으며 Boese 등 (2006)에 의해 추천된 신뢰구간과 비교하여 간편성, 포함확률의 근사성 및 구간의 길이에 있어 평균적으로 매우 우수하였다.

3. 이중표본추출에서 p의 신뢰구간

앞 절에서 살펴본 바와 같이 이중표본추출에서 스코어(score) 등 우도함수에 기초한 구간추정 방법은 장애모수 때문에 식 (2.7)과 같이 매우 복잡한 식으로 유도될 수 밖에 없으며, 이에 따라 매우 방대한 계산이 필연적으로 요구된다. 혹자는 컴퓨터의 발달과 더불어 현재의 통계적 방법에서 계산은 문제가 되지 않는다고 주장하기도 한다. 그러나 Agresti와 Coull (1998) 및 Brwon 등 (2001) 많은 문헌에서 주장한 바와 같이 구간추정 문제에서 계산의 간편성은 실용성과 직결되는 것으로 구간추정에서 매우 중요한 요인임을 간과할 수 없다. 즉, 지금까지 비율의 구간추정 문제에 있어 1) Clopper-Pearson exact 신뢰구간, 2) arcsine 신뢰구간, 3) logit 신뢰구간, 4) Jefferys 사전분포에 의한 신뢰구간, 5) Wilson 신뢰구간, 6) Agresti-Coull 신뢰구간, 7) Edgeworth 확장 등에 의한 신뢰구간 등 많은 구간추정방법이 Wald 신뢰구간의 대안으로 연구되어 왔으나 현재까지도 심각한 오류를 갖고 있는 Wald 신뢰구간이 가장 널리 사용되고 있는 이유도 Wald 신뢰구간이 다른 여러 신뢰구간과 비교하여 상대적으로 이해하기 쉬우며 간편하게 구할 수 있기 때문이라고 한다.

이러한 점들을 고려하면 이중표본추출에서는 Wald 신뢰구간과 같은 간편성이 확보된 Agresti-Coull 유형이 신뢰구간이 거의 유일한 대안이 될 수 밖에 없다. 따라서 본 연구에서는 모형 (2.1)에서 포함확률의 근사성 및 구간의 길이를 토대로 Wald 신뢰구간과 비교하여 Agresti-Coull 유형의 신뢰구간에 대해 효율성을 검토하기로 한다.

3.1. Wald 신뢰구간

Wald 신뢰구간은 식 (2.2)과 (2.5)을 이용하여

$$\mathrm{CI}_W(\mathbf{n}, x, y) = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{\mathrm{Var}}(\hat{p})} \tag{3.1}$$

와 같이 구할 수 있다. 여기서 $\mathbf{n}$은 $(n_{00}, n_{01}, n_{10}, n_{11})$를 나타낸다. 그러나 $n_{\cdot 0} = 0$ 또는 $n_{\cdot 1} = 0$이 되면 식 (2.2)의 $\hat{p}$가 정의되지 않는다. 이에 따라 $\widehat{\mathrm{Var}}(\hat{p})$ 역시 구할 수 없게 된다. 즉, Wald 신뢰구간이 존재하지 않는 경우가 확률 $\Pr[n_{\cdot 0} = 0] + \Pr[n_{\cdot 1} = 0] = (1-\pi)^n + \pi^n$으로 발생하게 된다. 이 확률은 n이 큰 경우 무시될 수도 있으나 거짓-양성 및 거짓-음성 오류율은 그다지 크지 않은 경우가 많기 때문

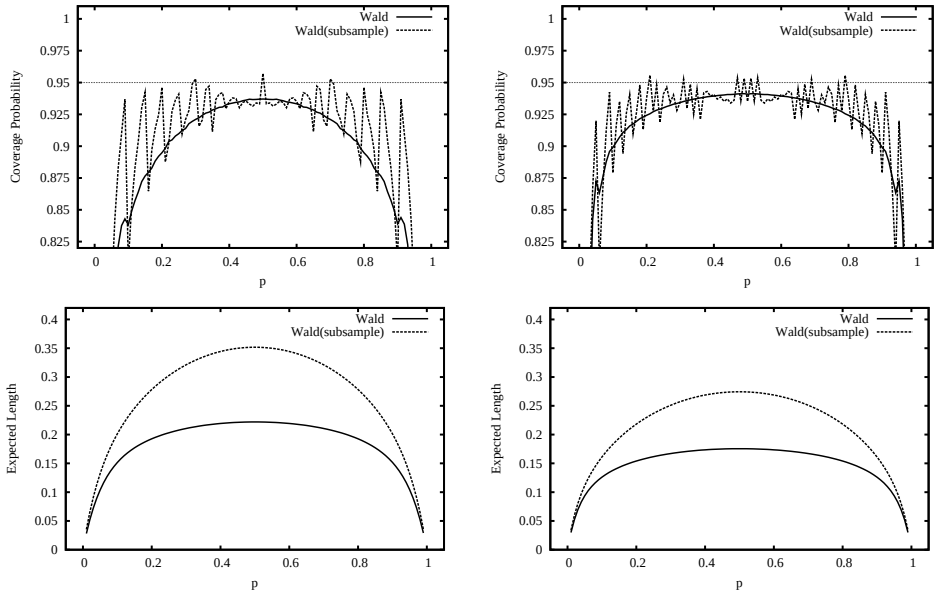

그림 **3.1.** $\theta = \phi = 0.1, n = 30$(좌측), 50(우측), $N = 300$(좌측), 500(우측) 일 때, 95% Wald 신뢰구간의 포함확률(위)과 기대길이(아래)

에 n이 크지 않을 경우 무시될 수 없는 확률을 갖을 수 있다. 예를 들어 $n = 30, \pi = 0.02$이면 $\hat{p}$가 정의되지 않을 확률은 0.5에 가깝다.

$n_{\cdot 0} = 0$ 또는 $n_{\cdot 1} = 0$인 경우 Tenenbein (1970)은 $Z = X + n_{\cdot 1}$으로 부터 $n_{\cdot 1} = nZ/N, n_{\cdot 0} = n - n_{\cdot 1}$과 같이 $n_{\cdot 0}$과 $n_{\cdot 1}$을 재정의하여 $\hat{p}$를 구하였다. 이러한 정의에 따르면 $n_{\cdot 1} = 0, X = 0$ 또는 $n_{\cdot 0} = 0, Y = 0$일 경우 같은 문제가 발생할 수도 있으나, N은 n과는 달리 비교적 큰 값을 갖는 것이 일반적이므로 현실적으로는 문제가 되지 않는다고 할 수 있다.

추정량이 존재하지 않을 확률 문제와는 달리 Wald 신뢰구간의 보다 본질적인 문제는 (2.5)의 근사 분산식은 실제 분산을 과소추정(underestimate)하게 만든다는 것이다. 대표본에서는 과소추정이 크게 문제가 되지 않을 수 있으나, 소표본에서는 식 (3.1)의 포함확률이 명목 신뢰수준에 훨씬 못미치는 결과를 가져오게 된다. 즉, Tenenbein (1970)은 1차 테일러 전개(Taylor expansion)만을 이용하여 식 (2.5)를 유도하였으나 1차 전개에서 생략된 항들의 누적효과 때문에 소표본에서는 과소추정 문제가 심각할 수 있다. 이러한 과소추정 문제의 효과는 그림 3.1에서 확인할 수 있다.

그림 3.1에는 $\theta = \phi = 0.1$을 가정하였을 때 (n, N)이 각각 $(30, 300)$ 및 $(50, 500)$에서 Wald 신뢰구간의 포함확률

$$C_{CI}(p; \theta, \phi) = \sum_{\mathbf{n}} \sum_{x,y} I(p \in \mathrm{CI}(\mathbf{n}, x, y)) L(p, \theta, \phi; \mathbf{n}, x, y) \tag{3.2}$$

과 기대길이

$$E_{CI}(p, \theta, \phi) = \sum_{\mathbf{n}} \sum_{x,y} \mathrm{Len}(\mathrm{CI}(\mathbf{n}, x, y)) L(p, \theta, \phi; \mathbf{n}, x, y) \tag{3.3}$$

를 나타낸 것이다. 식 (3.2)에서 I는 일반적인 지시함수(indicator function)를 나타내고, 식 (3.3)의 Len은 신뢰구간의 길이를 나타내는 함수이다. 그림에서 점선은 비교를 위해 부표본만으로 구한 Wald

신뢰구간의 포함확률과 기대길이를 나타낸 것이다. 그림에서 간편검사만 실시된 표본들 때문에 기대길이를 줄일 수 있었으나 포함확률의 근사성에는 거의 영향을 미치지 못하고 있는 것을 볼 수 있다. 즉, 식 (3.1) 신뢰구간의 포함확률이 명목신뢰수준에 근접하려면 전체 표본크기 N보다는 부표본의 크기 n이 증가되어야 한다는 것을 알 수 있다.

3.2. Agresti-Coull 신뢰구간

Agresti와 Coull (1998)은 $X \sim B(n,p)$라고 할 때 모비율 p의 신뢰구간을 구하기 위해 가상의 4개 관찰값(2개의 성공과 2개의 실패)를 더하고 Wald 식에 의해 신뢰구간을 구하였다. 즉, 이항분포에서 Agresti-Coull 신뢰구간은 $\tilde{p} = (X+2)/(n+4)$라고 할 때

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} \tag{3.4}$$

와 같이 주어진다. 그들은 식 (3.4)의 이론적 근거를 베이지안 방법론에 두고 있었고, 가상 관찰값의 수가 4개인 이유에 대해 Wilson 신뢰구간과 관련이 있음을 언급하였다. 왜 하필 4개인지에 대한 이론적 근거는 명확하지 않으나 앞 절에서 언급된 여러 연구에서 4개의 가상 관찰값이 매우 잘 작동을 하고 있음을 볼 수 있다. 특히, Lee (2007)에서는 거짓-양성 오류 이중표본추출 모형에서 Agresti-Coull 방법이 Boese 등 (2006)에서 추천된 식 (2.7) 보다 포함확률의 근사성 및 신뢰구간의 기대길이에서 우수하다는 것을 보이고 있다.

식 (2.1)은 4개의 범주를 갖는 다항분포와 이항분포의 결합분포로서 각 분포에 m_1과 m_2개의 가상 관찰값이 존재한다고 하자. 따라서 총 관찰값의 수는 $\tilde{N} = N + m_1 + m_2$개이다. 이때 $\tilde{n}_{tf} = n_{tf} + m_1/4$, $\tilde{X} = X + m_2/2$, $\tilde{Y} = Y + m_2/2$와 같이 정의된다. Agrest-Coull 유형신뢰구간은 식 (2.2)–(2.5)에 새로운 관찰값의 특성을 적용하여 구해진다. 이를

$$\mathrm{CI}_{Am}(\mathbf{n}, x, y) = \tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\widetilde{\mathrm{Var}}(\tilde{p})} \tag{3.5}$$

으로 표현하자. 여기서 m은 추가된 가상관찰값의 수, 즉 $m_1 + m_2$를 나타낸다.

새로운 관찰값에 의해 구해진 추정량은 일종의 축소추정량(shrinkage estimator)을 활용한 것으로 볼 수 있다. 예를 들어 식 (2.2)에서 $\pi_{1|1} = \Pr[T_1 = 1 | F_i = 1] = (1-\theta)p/\pi$의 우도추정량인 $n_{11}/n_{\cdot 1} = \hat{\pi}_{1|1}$과 $\tilde{n}_{11}/\tilde{n}_{\cdot 1} = \tilde{\pi}_{1|1}$는

$$\tilde{\pi}_{1|1} = \hat{\pi}_{1|1} + \frac{m_1/2}{n_{\cdot 1} + m_1/2} \left(\frac{1}{2} - \hat{\pi}_{1|1} \right) \tag{3.6}$$

와 같은 관계를 갖는다. 또, $\hat{\pi} = (X + n_{\cdot 1})/N$과 $\tilde{\pi} = (\tilde{X} + \tilde{n}_{\cdot 1})/\tilde{N}$의 관계도

$$\tilde{\pi} = \hat{\pi} + \frac{m_1 + m_2}{\tilde{N}} \left(\frac{1}{2} - \hat{\pi} \right) \tag{3.7}$$

와 같이 나타낼 수 있고, 식 (2.2)–(2.5)에 나타난 다른 우도추정량도 해당되는 Agresti-Coull 유형의 추정량과 비슷한 관계를 갖는다. 즉, Agresti-Coull 유형의 추정량은 표본비율과 1/2사이에 위치한 일종의 축소추정량이다.

식 (3.6) 또는 (3.7)과 같은 축소추정량은 평균제곱오차(mean squared error)의 측면에서 우도추정량보다 우수한 것으로 알려져 있으므로 결국 Agresti-Coull 유형의 신뢰구간은 평균제곱오차가 적은 축소추정량을 활용한 것이라고 할 수 있다. 한편 대표본에서 우도추정량의 정규성을 이용하면 이중표본추출 모형에서 (3.5)와 같은 정규근사에 의한 Agresti-Coull 신뢰구간에 대한 이론적 배경을 다음과 같이 설명할 수 있다.

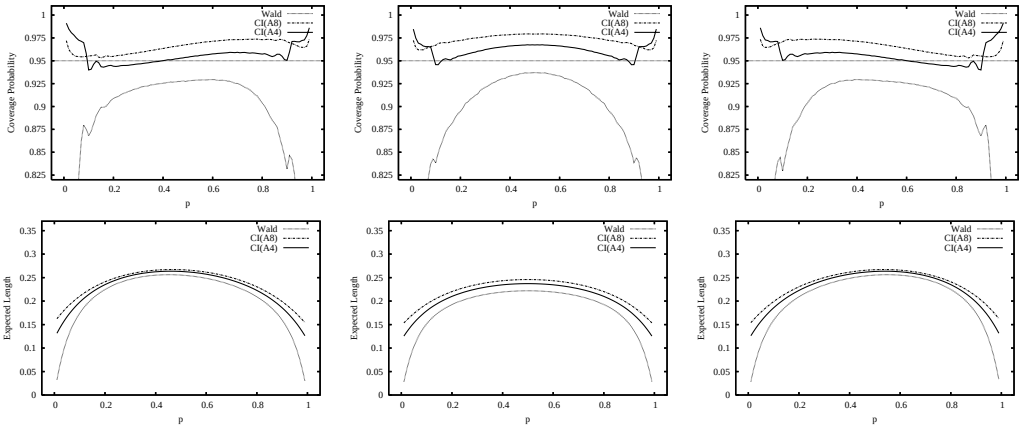

그림 **4.1.** $N = 300, n = 30, (\theta, \phi) = \{(0.2, 0.1), (0.1, 0.1), (0.1, 0.2)\}$일 때, 95% Wald와 Agresti-Coull 신뢰구간의 포함확률(위)과 기대길이(아래)

정리 3.1 $n \to \infty, N \to \infty$이고, $n/N \to f > 0$이면

$$\frac{\tilde{p} - p}{\sqrt{\widetilde{\mathrm{Var}}(\tilde{p})}} \xrightarrow{d} N(0, 1)$$

을 만족한다.

$\tilde{p}$와 $\hat{p}$의 정규분포로의 수렴속도는 동일하므로 이론적으로 대표본에서는 식 (3.5)는 (3.1)과 차이가 없다. 따라서 본 연구에서는 소표본에서 두 식에 의한 신뢰구간들의 특성을 살펴보기로 한다.

4. 소표본에서 신뢰구간들의 비교

Agrest와 Coull (1998)은 이항확률의 신뢰구간을 구하는 문제에서 4개의 가상 관찰값을 추가하여 성공적으로 신뢰구간을 구축하였고, Agresti와 Caffo (2000)에서도 4개의 가상 관찰값이 Wald 신뢰구간을 획기적으로 개선할 수 있었다. 특히, Agresti와 Caffo (2000)는 독립적인 2개의 이항분포에서 비율차에 대한 신뢰구간의 구축 문제에서 가장 적절한 가상 관찰값의 개수를 실험을 통해 구한 결과 4개의 가상 관찰값이 최적에 가깝다고 결론을 내렸다.

식 (2.1)은 다항분포와 이항분포의 결합분포이므로 비록 비율차에 대한 신뢰구간은 아니지만 Agresti와 Caffo (2000)에서 다루어진 모형과 유사하다고 할 수 있다. 따라서 다항분포와 이항분포에 각각 2개씩의 가상 관찰값을 부여한 Agresti-Coull 신뢰구간(CI_{A4})을 고려하기로 한다. 또한 각 분포에 4개씩 가상 관찰값이 추가된 모형(CI_{A8})도 비교를 위해 실험에 포함하였다. 그 밖에 몇 개의 다른 수의 가상 관찰값도 고려되었으나 생략되었다.

그림 4.1은 $N = 300, n = 30$일 때, θ와 ϕ의 값을 변화시키면서 구한 세 신뢰구간의 포함확률과 기대길이를 나타낸 것이다. 일반적으로 θ와 ϕ의 값은 그다지 큰 값이 아니기 때문에 각각 0.1 또는 0.2를 가정하였으나 그 밖의 값에 대해서도 세 신뢰구간의 대략적인 특징은 그림 4.1과 유사하다. 그림에서 보듯이 Wald 신뢰구간의 포함확률은 명목 신뢰수준에 미치지 못하고 있어 근사성이 좋지 않았으며 특히 p가 0 또는 1에 가까운 경우 즉, 분포가 한쪽으로 치우치면 신뢰구간으로서의 기능을 거의 하지 못한다고 볼 수 있다. p가 0.5 근처에서는 기대길이에 있어 Agresti-Coull 유형의 신뢰구간과 비교하여 큰 차

표 **4.1.** 95% 신뢰구간들의 평균 포함확률, 평균절대오차(Mean Absolute Error) 및 기대길이

				평균 포함확률 (MAE)						평균 기대길이		
θ	ϕ	N	n	CI_W		CI_{A8}		CI_{A4}		CI_W	CI_{A8}	CI_{A4}
0.1	0.1	200	20	0.821	(0.129)	0.981	(0.031)	0.971	(0.021)	0.213	0.276	0.259
			40	0.895	(0.056)	0.965	(0.015)	0.958	(0.008)	0.174	0.195	0.188
		300	30	0.865	(0.085)	0.973	(0.023)	0.962	(0.012)	0.184	0.220	0.208
			60	0.914	(0.036)	0.960	(0.010)	0.955	(0.005)	0.145	0.157	0.152
		400	40	0.889	(0.061)	0.968	(0.018)	0.958	(0.009)	0.163	0.187	0.178
			80	0.924	(0.026)	0.958	(0.008)	0.954	(0.004)	0.127	0.134	0.131
	0.2	200	20	0.830	(0.120)	0.970	(0.020)	0.961	(0.014)	0.243	0.292	0.280
			40	0.894	(0.056)	0.960	(0.010)	0.955	(0.006)	0.191	0.208	0.202
		300	30	0.870	(0.080)	0.965	(0.015)	0.956	(0.009)	0.208	0.235	0.227
			60	0.913	(0.037)	0.958	(0.008)	0.953	(0.004)	0.159	0.168	0.165
		400	40	0.891	(0.059)	0.962	(0.012)	0.954	(0.007)	0.184	0.202	0.196
			80	0.923	(0.027)	0.956	(0.006)	0.952	(0.003)	0.139	0.145	0.142
0.2	0.1	200	20	0.827	(0.123)	0.970	(0.020)	0.961	(0.014)	0.243	0.292	0.280
			40	0.894	(0.056)	0.960	(0.010)	0.955	(0.006)	0.191	0.208	0.202
		300	30	0.869	(0.081)	0.965	(0.015)	0.956	(0.009)	0.208	0.235	0.227
			60	0.913	(0.037)	0.958	(0.008)	0.953	(0.004)	0.159	0.168	0.165
		400	40	0.891	(0.059)	0.962	(0.012)	0.954	(0.007)	0.184	0.202	0.196
			80	0.923	(0.027)	0.956	(0.006)	0.952	(0.003)	0.139	0.145	0.142
	0.2	200	20	0.834	(0.116)	0.963	(0.013)	0.955	(0.008)	0.268	0.305	0.297
			40	0.895	(0.056)	0.958	(0.008)	0.953	(0.006)	0.206	0.219	0.215
		300	30	0.873	(0.077)	0.960	(0.010)	0.953	(0.007)	0.228	0.249	0.243
			60	0.913	(0.037)	0.956	(0.006)	0.952	(0.005)	0.171	0.178	0.175
		400	40	0.893	(0.057)	0.958	(0.008)	0.952	(0.006)	0.201	0.215	0.210
			80	0.923	(0.027)	0.955	(0.005)	0.952	(0.004)	0.149	0.154	0.152

이를 보이지 않았음에도 불구하고 포함확률의 값은 많은 차이가 있다. 이러한 사실은 신뢰구간의 위치를 결정하는 우도추정량이 구간추정에서는 그다지 효율적인 추정량이 아니라는 것을 의미한다. 이와 비교하여 두 Agresti-Coull 신뢰구간은 포함확률은 명목 신뢰수준을 상회하는 보수적인 신뢰구간으로 나타났으며, 분포가 한 쪽으로 치우친 경우에 보수성이 더욱 두드러진다. 이러한 특징은 Agresti-Coull 유형의 신뢰구간의 일반적인 특징으로 Agresti와 Coull (1998) 및 Agresti와 Caffo (2000)에서도 이러한 특징을 볼 수 있다.

비록 CI_{A4}와 CI_{A8}이 보수적이기는 하지만 포함확률이 Wald 신뢰구간보다 명목 신뢰수준에 근접하고 있어 포함확률의 근사성에서 Wald 신뢰구간보다 월등하므로 구간추정량으로서 Agresti-Coull 방법은 이중표본추출에서도 효능을 발휘하고 있다고 할 수 있다. 한편 Agresti-Coull 유형의 신뢰구간들을 비교하면 4개의 가상관찰값을 이용한 CI_{A4}가 CI_{A8}보다 덜 보수적이어서 포함확률의 근사성이 좋았고, 기대길이도 더 짧은 것을 살펴 볼 수 있다. 따라서 CI_{A4}가 p의 신뢰구간으로서 더 우수하다는 결론을 내릴 수 있다. 이러한 결론은 표 4.1에서 더욱 확실히 뒷받침될 수 있다. 즉, 표에서는 $p = 0.01, \ldots, 0.99$에서 구한 포함확률의 평균과 포함확률과 명목신뢰수준의 절대차이의 평균 즉, 평균절대오차 및 기대길이를 수록한 것으로 표에 나타난 모든 모수값에서 CI_{A4}의 포함확률이 명목 신뢰수준에 가장 근접하였고, CI_{A4}가 CI_{A8}보다 신뢰구간의 효율성을 나타내는 또다른 지표인 기대길이가 일률적으로 작은 값을 갖고 있다.

표 **5.1.** 네델란드 사회보장 집행오류 데이터

단계	감독기관 검사	회계감사관 검사	
		0	1
1	0	50	1
	1	0	2
2		433	14

표 **5.2.** 집행오류 비율의 95% 구간추정 결과

신뢰구간	하한	상한	구간길이
CI_W	0 (-0.003)	0.0484	0.0484
CI_{A4}	0 (-0.004)	0.0701	0.0701

표 4.1에서 모든 신뢰구간이 표본크기 특히 n이 커짐에 따라 표함확률의 근사성이 점차 개선되고 있는 것을 볼 수 있다. 그러나 Wald 신뢰구간은 표본크기 n이 비교적 큰 값이라고 할 수 있는 80인 경우도 근사성이 그다지 좋지 않으므로 다른 신뢰구간과 비교하여 개선속도가 많이 느린 것을 알 수 있다. 따라서 n이 아주 큰 경우가 아니라면 Wald 신뢰구간에 의한 추론은 매우 주의하여야 할 것이다.

5. 응용 예제

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 유명하여 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 표 5.1과 같다.

데이터에서 우도추정과 4개의 가상 관찰값을 추가하여 구한 추정값은 각각 $\hat{p} = 0.0227$, $\widehat{\mathrm{Var}}(\hat{p}) = 0.0001718$과 $\tilde{p} = 0.0330$, $\widetilde{\mathrm{Var}}(\tilde{p}) = 0.0003577$과 같이 구하여져 표 5.2과 같은 신뢰구간을 구할 수 있었다. 표 5.2에서 괄호안의 값은 구간추정식 (3.1)과 (3.5)에 의해 구한 값이나 음의 값은 모수공간에 포함되지 않으므로 절사하여 구간의 하한을 0으로 설정한 것이다.

특이한 것은 두 추정에서 장애모수인 θ와 ϕ는 각각 $\hat{\theta} = 0.000$, $\hat{\phi} = 0.0116$과 $\tilde{\theta} = 0.2859$, $\tilde{\phi} = 0.0146$으로 구하여져 다른 모수와는 달리 θ의 추정에서 두 추정값이 상당한 차이를 보이고 있다는 점이다. 우도추정에서 θ의 추정값이 0이 된 이유는 전적으로 $n_{01} = 0$인 때문이다. 즉, 추정값이 부표본(subsample)에만 의존하게 되었기 때문이다. 이와는 달리 $\tilde{\theta}$는 전체 표본에 의해 조정된 값이라고 볼 수 있다. 실제로 범주 $(0,0)$에서 하나의 관찰값을 $(1,0)$으로 옮기게 되면, 다시 말하면 $n_{00} = 49, n_{10} = 1$이 되면 θ의 우도추정값은 0.4602로 변하게 된다. 따라서 θ의 우도추정량은 매우 분산이 크다는 것을 알 수 있다. 이 경우 $\tilde{\theta} = 0.5457$로 변하게 되므로 변동의 폭은 상대적으로 우도추정량보다 적다는 것을 알 수 있다.

점추정의 효율성과는 별도로 실제 θ의 값이 우도추정값과 같이 매우 작은 값이라고 하여도 Wald 신뢰구간은 포함확률의 근사성에서 매우 좋지 않았다. 예를 들어 장애모수를 우도추정값에 아주 가깝게 $\theta = 0.01$, $\phi = 0.012$로 설정한 경우에도 $p = 0.02$ 또는 0.03에서 Wald 신뢰구간의 포함확률은 각각

0.593과 0.736이었으며, Agresti-Coull 신뢰구간의 경우 포함확률은 모두 0.998이었다. 비록 Agresti-Coull 신뢰구간이 매우 보수적이기는 하지만 두 구간의 길이 차이는 0.022에 불과하므로 구간 길이에서 큰 차이가 있다고 보기는 어렵다. 그러므로 네델란드 사회보장 집행오류 데이터의 분석에서 Agresti-Coull 신뢰구간은 Wald 신뢰구간보다 타당성이 높다고 할 수 있다.

참고문헌

0이승천 (2006). 독립표본에서 두 모비율 차이에 대한 가중 Polya 사후분포 신뢰구간, <응용통계연구>, **19**, 171–181.

0이승천 (2007). 베이지안 접근에 의한 모비율 선형함수의 신뢰구간, <응용통계연구>, **20**, 257–266.

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

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, *American Statistician,* **54**, 280–288.

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

Boese, D. H., Young, D. M. and Stamey, J. D. (2006). Confidence intervals for a binomial parameter based on binary data subject to false-positive misclassification, *Computational Statistics and Data Analysis,* **50**, 3369–3385.

Braunstein, G. (2002). False-positive serum human chronic gonadotropin results: causes, characteristics, and recognition, *American Journal of Obstetrics & Genecology,* **187**, 217–224.

Bross, I. (1954). Misclassification in tables, *Biomometrics,* **10**, 478–486.

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

Kazemi, N., Dennien, B. and Dan, A. (2001). Mistaken identity: A case of false positive on CT angiography, *Journal of Clinical Neuroscience,* **9**, 464–466.

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

Lee, S.-C. (2007). An improved confidence interval for the population proportion in a double sampling scheme subject to false-positive misclassification, *Journal of the Korean Statistical Society,* **36**, 275–284.

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

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

Swaen, V. M., Teggerler, O. and Amelsvoort, L. (2001). False positive outcomes and design characteristics in occupational cancer epidemiology studies, *International Journal of Epidemiology,* **30**, 948–955.

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

Tenenbein, A. (1971). A double sampling scheme for estimating from binomial data with misclassifications: sample size determination, *Biometrics,* **27**, 935–944.

Tenenbein, A. (1972). A double sampling scheme for estimating from multinomial data with application to sampling inspection, *Technometrics,* **14**, 187–202.

York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995). Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty, *Applied. Statistics,* **44**, 227–242.

Interval Estimation of Population Proportion in a Double Sampling Scheme

Seung-Chun Lee[1] · Byong Su Choi[2]

[1]Department of Statistics, Hashin University
[2]Department of Multimedia Engineering, Hansung University

Abstract

The double sampling scheme is effective in reducing the sampling cost. However, the doubly sampled data is contaminated by two types of error, namely false-positive and false-negative errors. These would make the statistical analysis more difficult, and it would require more sophisticate analysis tools. For instance, the Wald method for the interval estimation of a proportion would not work well. In fact, it is well known that the Wald confidence interval behaves very poorly in many sampling schemes. In this note, the property of the Wald interval is investigated in terms of the coverage probability and the expected width. An alternative confidence interval based on the Agresti-Coull's approach is recommended.

Keywords: False-positive error, false-negative error, Wald confidence interval, Agresti-Coull confidence interval, coverage probability.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2008-313-C00131).

The second author was financially supported by Hansung University in the year of 2008.

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