

이항 비율의 가중 POLYA POSTERIOR 구간추정*

이승천¹⁾

요약

최근 여러 학자들에 의해 이항 비율의 구간 추정에 많이 사용되고 있는 Wald 신뢰구간의 문제점이 재조명되고 있고, 이에 대한 대안으로 이항 비율의 새로운 신뢰구간들이 발표되고 있다. 본 논문에서는 가중 Polya posterior를 이용한 베이지안 구간추정을 구하였다. 이 구간추정은 이항분포의 공액분포인 베타 사전분포에서 구한 전통적인 베이지안 구간추정과 같으나 추정의 편의를 위하여 정규근사에 의한 신뢰구간을 구할 때, 표본크기가 크면 실제로 Argresti와 Coull (1998)의 신뢰구간과도 일치하였다. 또 새로운 신뢰구간은 표본크기가 작은 경우와 비율이 극히 작은 경우에도 매우 유용한 신뢰구간이 된다는 것을 살펴 보았다.

주요용어: 이항 비율, 가중 Polya posterior, Wald 신뢰구간, Agresti-Coull 신뢰구간, Wilson 신뢰구간

1. 서론

이항 비율의 신뢰구간 추정은 고전적이면서도 매우 중요한 문제로 정규근사에 의한 Wald 신뢰구간이 일반적으로 많이 사용되고 있다. Wald의 신뢰구간이 많이 사용되고 있는 이유의 하나는 간단한 식에 의해 쉽게 계산할 수 있다는 장점에 기인하는 바가 크다고 하겠다. 그러나 Ghosh (1979), Blyth와 Still (1983) 등 많은 연구에서 Wald의 신뢰구간의 포함확률(coverage probability)은 신뢰수준과 상당한 차이를 가지고 있으며 특히, 표본크기가 작은 경우와 모비율이 작은 경우 문제가 심각하다는 점을 지적하였다. 이에 Agresti와 Coull (1998), Agresti와 Caffo (2000), Brown 등 (2001, 2002)은 이항 비율의 새로운 신뢰구간을 발표하였다. Agresti와 Coull의 신뢰구간은 기존의 Wald 신뢰구간을 간단히 수정한 것으로 컴퓨터의 도움없이도 간편히 계산할 수 있으며 효율성에 있어서도 매우 우수한 것으로 인정되고 있다.

한편, 이항 비율의 베이지안 구간추정은 Jefferys 사전분포를 이용하여 구한 최고사후밀도구간이 많이 사용되고 있다. Brown 등 (2001)에 의하면 베이지안 신뢰구간(credible interval)은 표본크기가 작은 경우는 ($n \leq 40$) 물론 표본크기가 큰 경우에서도 Wilson (1927) 신뢰구간과 함께 매우 우수하다고 한다. 그러나 Jefferys 최고사후밀도구간은 임계값의 계산이 용이하지 않다는 문제를 갖는다. 비록 Brown (2001)이 임계값에 대한 근사식을 제시

* 이 논문은 2005년도 한신대학교 학술연구비 지원에 의하여 연구되었음

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수

E-mail: seung@hanshin.ac.kr

하였으나 그리 간단치 않은 계산을 요구하므로 대개의 경우 컴퓨터의 도움을 필요로 한다고 하겠다.

Ghosh와 Meeden (1998)에 의하면 유한 모집단 표본에서 표본평균과 같이 통계분석에서 많이 사용되고 있는 통계량들이 단계별 베이스 (stepwise Bayes) 방법에 의한 베이지안 통계량임을 보여 빈도학파의 많은 추정량들이 허용성(admissibility)을 갖는다는 사실을 밝혀내었다. 그들이 사용한 주요한 도구는 무정보 단계별 사전분포에 의한 Polya posterior이다. 특히, Meeden (1999)은 가중 Polya posterior를 이용하여 왜도가 심한 유한모집단에서 표본 평균에 대한 구간추정을 하였는데, 여기서 적용된 단계별 베이스 방법은 이항 비율의 구간추정에 있어서도 적용이 가능하다. 일반적으로 비율 구간추정에서의 문제는 비율 p 가 0 또는 1에 가까운 경우 즉, 왜도가 심한 경우 발생하게 되므로 비율의 구간추정에서 가중 Polya posterior는 매우 유용할 것으로 예상할 수 있다.

2. 가중 Polya posterior

이 절의 기호는 Ghosh와 Meeden (1998)을 따르기로 한다. 즉, $U = \{y_1, y_2, \dots, y_N\}$ 를 N 개의 요소로 이루어진 유한 모집단이라고 하고, n 을 표본크기라고 하자. 또 s 를 n 개의 요소를 갖는 $\{1, 2, \dots, N\}$ 의 부분집합으로 이를 표본이라 한다. 즉, $s = \{i_1, i_2, \dots, i_n\}$ 이면 단순임의 표본추출에 의해 $\{y_{i_1}, y_{i_2}, \dots, y_{i_n}\}$ 이 표본 추출되었다는 것을 의미한다. 또 $z_s = \{z_{i_1}, z_{i_2}, \dots, z_{i_n}\}$ 는 주어진 표본에서 측정된 관측값을 나타내기로서 하며, 비율의 추론 문제에 있어서 z_i 는 0 또는 1의 값을 갖게 된다.

Polya posterior는 주어진 데이터 (s, z_s) 에 의한 미관측된 값들에 대한 예측 결합분포(predictive joint distribution)으로 Feller (1968)에 잘 기술되어 있다. 즉, 항아리에 $z_{i_1}, z_{i_2}, \dots, z_{i_n}$ 의 값이 적혀진 n 공이 있는데, 여기서 하나의 공을 임의로 추출한다. 추출된 공에 적혀진 값을 새로운 공에 기입하여 추출된 공과 같이 다시 항아리에 넣는다. 그러므로 항아리에는 $n+1$ 개의 공이 있게 된다. 이러한 과정을 $N-n$ 번 반복하여 N 개의 공을 항아리에 채우게 된다. 이렇게 해서 얻어진 N 개의 값들은 주어진 표본에 의해 예측된 모집단으로 인식하게 되는 것으로 예측분포로서의 타당성 등 보다 자세한 내용은 Ghosh와 Meeden (1998)를 참조할 수 있다.

데이터 (s, z_s) 에서 0과 1의 값을 갖는 갯수를 각각 n_0, n_1 ($n_0 + n_1 = n$)이라 하자. Polya posterior는 모든 공이 추출된 확률이 같도록 하였으나 가중 Polya posterior는 공의 가중치를 달리 하여 얻어진 것이다. 즉, $\xi (> 0)$ 를 임의의 상수라고 할 때, 초기에 0과 1이 적혀진 공에 대해서는 각각 $1 + \xi/n_0$ 와 $1 + \xi/n_1$ 으로 가중치를 부여한다. 이 경우 총 가중치의 합은 $n + 2\xi$ 가 되어 첫번째 시도에서 1이 적혀진 공이 추출된 확률은 $(n_1 + \xi)/(n + 2\xi)$ 이 된다. 이후 새롭게 추가되는 공들에 대해서는 모두 1의 가중치를 갖는다. 즉, 관찰된 데이터에 보다 많은 가중치를 부여한 것이다.

U 가 무한모집단임을 가정하여 추출, 복원 과정을 무수히 반복한다고 가정하자. 이렇게 만들어진 예측 모집단에서의 1의 비율을 p 라 하면 de Finetti 정리를 이용하여 $p | (s, z_s)$ 의 극한분포가 $\text{Beta}(n_1 + \xi, n - n_1 + \xi)$ 이라는 것을 증명할 수 있다. 그러나 본 논문에서는 단

지 극한 분포의 평균과 분산만을 필요로 하므로 이 두 값만을 구하여 보기로 한다.

가중 Polya posterior에서 표본 (s, z_s) 가 주어졌을 때 $z_i, i \notin s$ 들은 교환가능한 변수(exchangeable variable)임이 알려져 있다. 그러므로 주어진 표본 (s, z_s) 에서 모평균 또는 모비율을 $p = \sum_{i=1}^N z_i/N$ 의 조건부 기대값에 대해 다음이 성립된다.

$$\begin{aligned} E(p | (s, z_s)) &= \frac{1}{N} \left\{ \sum_{i \in s} z_i + (N - n)E(z_1 | (s, z_s)) \right\} \\ &= \frac{1}{N} \left\{ n_1 + (N - n) \frac{n_1 + \xi}{n + 2\xi} \right\} = \hat{p} + \frac{2(N - n)\xi}{N(n + 2\xi)} \left(\frac{1}{2} - \hat{p} \right) \end{aligned} \quad (2.1)$$

한편 모비율의 조건부 분산은

$$\begin{aligned} \text{Var}(p | (s, z_s)) &= \frac{1}{N^2} \text{Var} \left(\sum_{j \notin s} z_j | (s, z_s) \right) \\ &= \frac{1}{N^2} \left\{ \sum_{j \notin s} \text{Var}(z_j | (s, z_s)) + \sum_{i \notin s \text{ and } j \notin s} \text{Cov}(z_i, z_j | (s, z_s)) \right\} \\ &= \frac{1}{N^2} \left\{ (N - n) \text{Var}(z_1 | (s, z_s)) + 2 \binom{N - n}{2} \text{Cov}(z_1, z_2 | (s, z_s)) \right\} \end{aligned} \quad (2.2)$$

와 같이 유도되는데 $\text{Var}(z_1 | (s, z_s))$ 와 $\text{Cov}(z_1, z_2 | (s, z_s))$ 는 각각

$$\begin{aligned} \text{Var}(z_1 | (s, z_s)) &= \frac{n_1 + \xi}{n + 2\xi} \left(1 - \frac{n_1 + \xi}{n + 2\xi} \right) \\ \text{Cov}(z_1, z_2 | (s, z_s)) &= E(z_1 z_2 | (s, z_s)) - E(z_1 | (s, z_s))^2 \\ &= E(z_1 E(z_2 | z_1, (s, z_s)) | (s, z_s)) - E(z_1 | (s, z_s))^2 \\ &= \frac{n_1 + \xi + 1}{n + 2\xi + 1} \frac{n_1 + \xi}{n + 2\xi} - \left(\frac{n_1 + \xi}{n + 2\xi} \right)^2 = \frac{1}{n + 2\xi + 1} \text{Var}(z_1 | (s, z_s)) \end{aligned}$$

이므로 (2.2)의 조건부 분산은

$$\text{Var}(p | (s, z_s)) = \frac{N - n}{N^2} \frac{N + 2\xi}{n + 2\xi + 1} \text{Var}(z_1 | (s, z_s)) \quad (2.3)$$

과 같이 구하여 진다. 단, 평균과 분산의 유도 과정에서 기호의 편의상 1과 2는 s 에 속하지 않는다고 가정하였고 $\hat{p} = n_1/n$ 으로 정의하였다.

이제 U 가 무한모집단임을 가정하여 $N = \infty$ 이면 추출을 무수히 반복 실행하면 모비율의 조건부 평균과 분산은 각각

$$E(p | (s, z_s)) \longrightarrow \hat{p} + \frac{2\xi}{n + 2\xi} \left(\frac{1}{2} - \hat{p} \right) = \frac{n_1 + \xi}{n + 2\xi} \quad (2.4)$$

$$\text{Var}(p | (s, z_s)) \longrightarrow \frac{1}{n + 2\xi + 1} \text{Var}(z_1 | (s, z_s)) = \frac{1}{n + 2\xi + 1} \frac{n_1 + \xi}{n + 2\xi} \left(1 - \frac{n_1 + \xi}{n + 2\xi} \right) \quad (2.5)$$

으로 수렴된다는 것을 알 수 있다. 이하에서는 1의 갯수 n_1 을 성공횟수 X 로 표기하기로 한다.

3. 이항 비율에 대한 Poylar posterior 근사 구간추정

무한 모집단에서 모비율 p 의 사후분포가 $\text{Beta}(X + \xi, n - X + \xi)$ 이라고 하였으므로 $100(1 - \xi)\%$ 베이지안 신뢰구간은 $\text{Beta}(\alpha; m_1, m_2)$ 를 $\text{Beta}(m_1, m_2)$ 의 α -분위수라고 할 때

$$[\text{Beta}(\alpha/2; X + \xi, n - X + \xi), \text{Beta}(1 - \alpha/2; X + \xi, n - X + \xi)]$$

와 같이 구할 수 있다. 이 결과는 이항분포의 공액분포족에 속하는 사전분포 $\text{Beta}(\xi, \xi)$ 를 이용한 베이지안 구간추정과 동일하며 특히, $\xi = 1/2$ 인 경우 Jefferys 사전분포 구간추정과 같다. 그러므로 앞에서 설명된 것과 같이 매우 효율적인 구간추정을 할 수 있다. 그러나 Brown 등 (2001)은 비율의 구간추정에서 추정식의 간편성에 대한 중요성을 언급하였으며 이러한 간편성 때문에 n 이 클 때 Agresti-Coull 신뢰구간을 추천하였다. 베이지안 신뢰구간은 효율성이 인정되고 있지만 임계점의 계산을 용이치 않은 단점을 갖고 있다. 비록 Brown 등 (2001)에 의해 임계점의 계산을 위한 근사식이 제시되었지만 간편성에서는 역시 문제를 갖게 되므로 여기에서는 베이지안 추정량의 BAN(Best Asymptotic normality) 특성에 의한 근사 신뢰구간을 고려하기로 한다. 즉, (2.4)과 (2.5)을 이용하여 실용성이 있는 근사 신뢰구간을 구하기로 한다.

(2.4)에 주어진 $\hat{p} + \frac{2\xi}{n+2\xi}(\frac{1}{2} - \hat{p})$ 는 표본비율과 $1/2$ 의 중간에 위치하는 일종의 축소추정량이라고 할 수 있다. 이와 같이 축소추정량을 이용한 구간추정량으로 Wilson (1927)의 신뢰구간

$$CI_W = \tilde{p} \pm \frac{z_{\alpha/2} n^{1/2}}{n + z_{\alpha/2}^2} \sqrt{\hat{p}(1 - \hat{p}) + \frac{z_{\alpha/2}^2}{4n}} \quad (3.1)$$

과 Agresti와 Coull (1998)의 신뢰구간

$$CI_{AC} = \tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + z_{\alpha/2}^2}} \quad (3.2)$$

이 있다. 여기서 $\tilde{p} = (X + z_{\alpha/2}^2/2)/(n + z_{\alpha/2}^2)$ 인데 $\xi = z_{\alpha/2}^2/2$ 일 때 (2.4)는 \tilde{p} 와 같다. (본래의 Agresti-Coull 신뢰구간에서는 $\alpha = 0.05$ 일 때, $\tilde{p} = (X + 2)/(n + 4)$ 으로 정의하였으나 여기에서는 Brown 등 (2001)의 정의에 따르기로 한다.) 이 경우 정규근사에 의한 가중 Polya posterior 신뢰구간은

$$CI_{PP} = \tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 1 + z_{\alpha/2}^2}} \quad (3.3)$$

와 같이 구하여진다. 그러므로 CI_{AC} 와 단지 $n + 1 + z_{\alpha/2}^2$ 와 $n + z_{\alpha/2}^2$ 의 차이만 있게 되며, n 이 큰 경우 두 신뢰구간은 사실상 같은 신뢰구간이라고 하겠다. Brown 등 (2001)에 의하면 표본크기가 40이상인 경우는 Wald 신뢰구간을 제외한 모든 구간들이 유효하다고 하였으므로 여기에서는 표본크기가 작은 경우 세 신뢰구간의 효율성을 비교하기로 한다.

Agresti-Coull의 신뢰구간은 모든 X 의 값에서 항상 Polya posterior 베이지안 구간은 물론 Wilson의 신뢰구간을 포함한다. 즉,

$$CI_W \subseteq CI_{AC}, \quad CI_{PP} \subseteq CI_{AC}$$

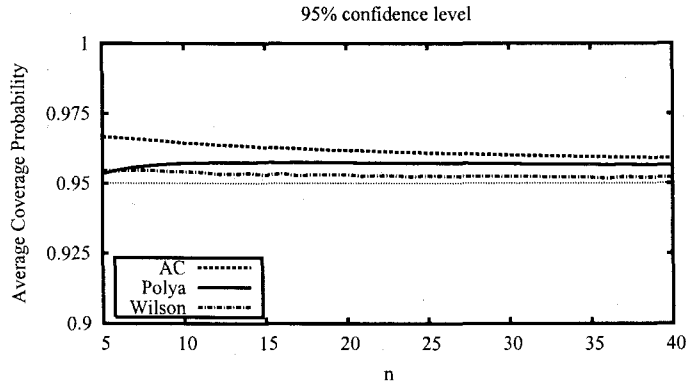


그림 3.1: 95% 신뢰수준에서 표본크기에 따른 Agresti-Coull, 가중 Polya posterior, Wilson 신뢰구간의 평균 포함확률

이 성립된다. 이로 인하여 CI_{AC} 의 포함확률(coverage probability)은 실제 신뢰수준보다 약간 큰 값을 갖는 경향이 있다. 이는 그림 3.1에서 확인할 수 있다. 그림 3.1는 표본크기 n 과 모비율 p 가 주어졌을 때 신뢰구간 CI 의 포함확률을 $C_{CI}(p, n)$ 라고 하면 $\int_0^1 C_{CI}(p, n) dp$ 를 표본크기에 따라 그린 것으로 표본크기에 따른 각 신뢰구간의 평균 포함확률을 나타낸 것이라고 할 수 있다. 그림의 상단에서부터 차례로 Agresti-Coull, 가중 Polya posterior, Wilson 신뢰구간들의 평균 포함확률을 표시하고 있다. 그림에서 Wilson의 신뢰구간이 Agresti-Coull 신뢰구간보다 평균적으로 적절한 평균포함확률을 갖고 있음을 알 수 있다. 결국 베이저의 입장에서는 Agresti와 Coull (1998)은 추정량의 분산을 과대추정하여 신뢰구간을 구한 것으로 생각할 수 있다. 한편 CI_{PP} 의 구간 길이는 Agresti-Coull의 구간 길이 보다 항상 작으므로 이러한 경향이 어느 정도 해소할 수 있으며 특히, 표본크기가 작은 경우 그 효과가 크다.

그림 3.1에서 세 개의 신뢰구간 중 Wilson의 신뢰구간은 평균적으로 가장 적절한 것으로 나타났지만 이 결과가 Wilson 신뢰구간이 가장 좋다는 것을 의미하는 것을 아니다. 그림 3.2는 각 표본 크기에서 $\int_0^1 |C_{CI}(p, n) - \alpha| dp$ 즉, 평균 절대오차를 나타낸 그림으로 대략 $n \leq 25$ 에서는 Polya posterior 신뢰구간의 평균 절대오차가 Wilson 신뢰구간보다 작은 값을 갖고 있음을 보여준다. 즉, 표본크기가 작은 경우 Wilson 구간의 포함확률은 신뢰수준을 중심으로 상하 변동이 상대적으로 Polya posterior 구간보다 크다는 것을 알 수 있다. 이는 그림 3.3에서도 확인할 수 있다.

그림 3.3에서 p 의 중앙 부분에서 두 신뢰구간의 포함확률은 큰 차이를 보이지 않지만 양 끝에서 차이를 보이고 있다. Wilson의 경우 신뢰수준 0.95를 중심으로 상하 변동을 하고 있고 특히 p 가 0 또는 1에 가까울 때 매우 큰 진폭을 보이고 있다. 즉, p 가 0 또는 1에 가까울 때 Wilson 신뢰구간의 포함확률은 신뢰수준과 비교하여 작은 값을 갖는 경우가 있다. Brown 등 (2001)은 이러한 경향을 Wilson 구간이 대못(spike)을 갖는다고 표현하였는데

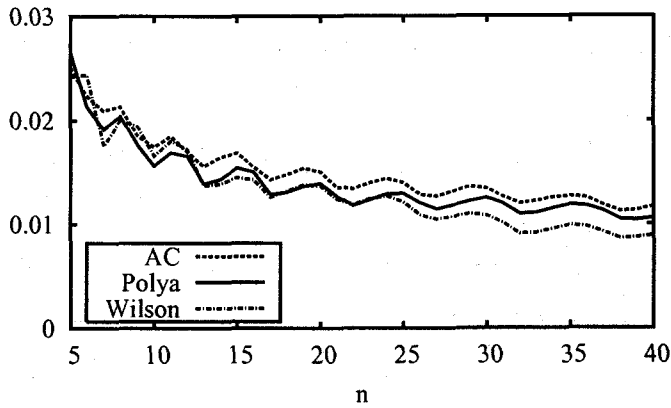


그림 3.2: Agresti-Coull, Poly posterior 베이지, Wilson 95% 신뢰구간들의 포함확률 평균 절대오차

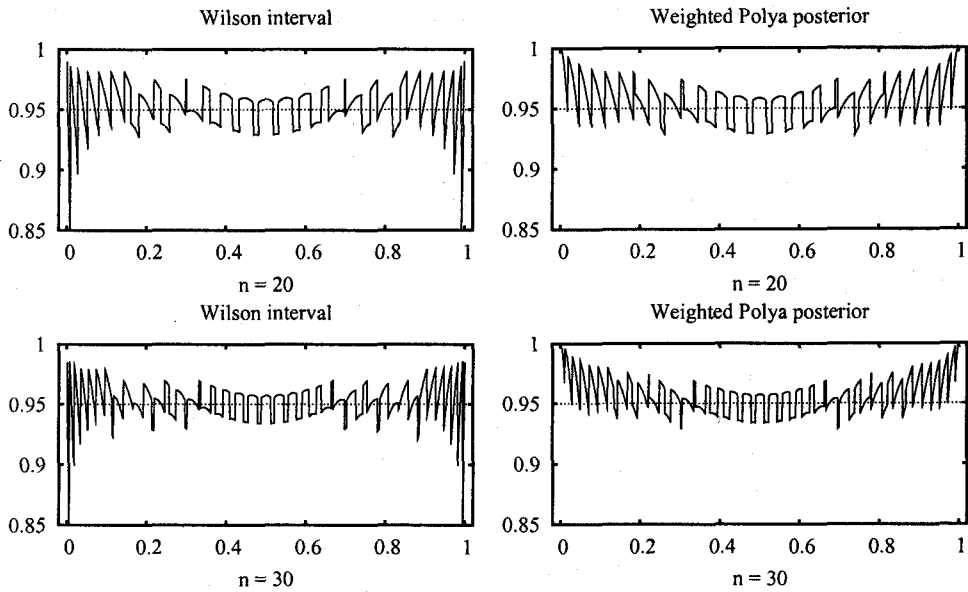


그림 3.3: 표본크기 $n = 20, 30$ 에서 Poly posterior와 Wilson 95% 신뢰구간의 포함확률

$5 \leq n \leq 40$ 에서 모두 Wilson 구간의 대못을 확인할 수 있었다. 표 3.1는 각 표본크기 n 에서 Wilson 구간의 가장 큰 대못을 나타낸 표이다.

한편 가중 Poly posterior 신뢰구간은 p 가 0과 1에 가까운 경우는 대체로 신뢰수준보

표 3.1: Wilson 95% 신뢰구간의 대못에서의 포함확률

n	5	10	15	20	25	30	35	40
p	0.0362	0.0178	0.0118	0.0088	0.0070	0.0059	0.0050	0.0044
포함확률	0.8316	0.8356	0.8369	0.8380	0.8389	0.8373	0.8391	0.8383

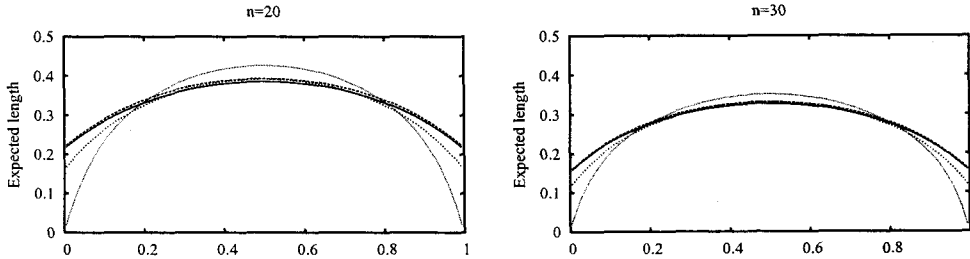


그림 3.4: 표본크기 $n = 20, 30$ 에서 95% 신뢰구간의 평균 길이 (왼쪽에 나타난 크기의 순서는 AC, Polya, Wilson, Wald이다.)

다 큰 값을 갖는 보수적인 신뢰구간을 형성한다. CI_{PP} 와 CI_W 를 비교하면 성공횟수 X 가 $n/2$ 에 가까운 경우에는 $CI_W \supset CI_{PP}$ 이지만 X 의 값이 $n/2$ 에서 멀어져 0 또는 n 에 가까워지면 $CI_W \subset CI_{PP}$ 이 된다. 즉, p 가 양 극값에 가까울 때 Polya posterior 베이지 구간은 Wilson의 구간보다 긴 신뢰구간을 형성하는 경향이 있다. 신뢰구간 CI의 신뢰수준이 $1 - \alpha$ 라는 것은 $\Pr[p \in CI] \geq 1 - \alpha$ 이므로 Polya posterior 신뢰구간의 신뢰수준은 Wilson 신뢰구간의 신뢰수준보다 정의에 부합된다고 볼 수 있다.

신뢰구간을 평가하는데 신뢰구간의 길이도 중요한 요소이므로 신뢰구간의 평균 길이를 비교하여야 한다. 이때 신뢰구간의 평균 길이는 x 에서 신뢰구간 CI의 하한과 상한을 각각 $L(x, n), U(x, n)$ 이라고 하면 CI의 평균 길이는

$$E_n(CI) = \sum_{x=0}^n (U(x, n) - L(x, n)) \binom{n}{x} p^x (1-p)^{n-x}$$

와 같이 정의된다.

그림 3.4은 신뢰수준 95%에서 신뢰구간들의 평균길이를 나타낸 것이다. 그림의 왼쪽과 오른쪽 부분에서의 크기 순서대로 나열하면 이미 언급한 바와 같이 Agresti-Coull, Polya posterior 베이지, Wilson, Wald의 순서대로 나타난다. 그러나 중앙부분에서는 Polya posterior 베이지 신뢰구간의 평균 길이가 Wilson 보다 아주 근소하게 작은 것을 볼 수 있다. 아주 극단적인 경우에도 가중 Polya posterior 신뢰구간의 평균길이는 Wilson과 비교하여 0.03 정도이므로 구간의 길이가 지나치게 길다고 하기는 어렵다.

4. 결론

현재까지 제안된 이항 비율의 신뢰구간들은 대체로 p 가 0.5에 가까우면 Wald 구간을 제외한 모든 신뢰구간의 포함확률이 신뢰수준과 큰 차이를 보이지 않으나 p 가 0 또는 1에 가까운 경우 각 신뢰구간의 포함확률이 다른 경우가 많다. 그러므로 실용적인 면에서 신뢰구간의 비교는 왜도가 심한 경우에는 비교되는 것이 바람직하다. 이러한 측면에서 가중 Polya posterior 베이즈 방법은 표본크기가 작고, 왜도가 큰 경우에 효과적인 신뢰구간을 구하여 주었다.

Brown 등 (2001)은 표본 크기가 작은 경우($n \leq 40$)에는 Wilson과 Jefferys 신뢰구간을 추천하였고, 표본 크기가 큰 경우 에서는 Wilson, Agresti-Coull 구간 등 Wald를 제외한 모든 신뢰구간이 사용될 수 있다고 하였다. 특히, Agresti-Coull 신뢰구간은 식 (3.2)에서 $\bar{n} = n + z_{\alpha/2}^2$ 이라고 하면 Wald와 신뢰구간과 같은 간단한 표현식을 갖게 되어 계산상으로도 매우 유리하다고 하였다. 가중 Polya posterior 베이즈 구간은 Agresti-Coull 신뢰구간을 간단히 수정하여 얻을 수 있는 것으로 계산의 간편성은 물론 표본 크기가 작고 왜도가 큰 경우에도 매우 효과적인 구간이 되므로 Wilson의 신뢰구간과 같이 표본크기에 관계 없이 추천할 만한 구간추정이라고 하겠다.

참고문헌

- 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, *The American Statistician*, **54**, 280-288.
- 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.
- Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions, *The Annals of Statistics*, **30**, 160-201.
- Feller, W. (1968). *An Introduction of Probability Theory and Its Applications*, volumn I, Wiley, New York.
- Ghosh, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter, *Journal of the American Statistical Association*, **74**, 894-900.
- Ghosh, M. and Meeden, G. D. (1998). *Bayesian Methods for Finite Population Sampling*, Chapman & Hall, London.
- 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.
- Wilson, E. B. (1927). Probable inference, the law of succession and statistical inference, *Journal of the American Statistical Association*, **22**, 209-212.

[2005년 2월 접수, 2005년 6월 채택]

Interval Estimation for a Binomial Proportion Based on Weighted Polya Posterior*

Seung-Chun Lee¹⁾

ABSTRACT

Recently the interval estimation of a binomial proportion is revisited in various literatures. This is mainly due to the erratic behavior of the coverage probability of the well-known Wald confidence interval. Various alternatives have been proposed. Among them, Agresti-Coull confidence interval has been recommended by Brown *et al.* (2001) with other confidence intervals for large sample, say $n \geq 40$. On the other hand, a noninformative Bayesian approach called Polya posterior often produces statistics with good frequentist's properties. In this note, an interval estimator is developed using weighted Polya posterior. The resulting interval estimator is essentially the Agresti-Coull confidence interval with some improved features. It is shown that the weighted Polya posterior produce an effective interval estimator for small sample size and a severely skewed binomial distribution.

Keywords: Binomial proportion, Weighted Polya posterior, Wald interval, Agresti-Coull interval, Wilson interval

* This Work was Supported by Hanshin University Research Grant in 2005.

1) Professor, Dept. of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do, 447-791
E-mail: seung@hanshin.ac.kr