

이항모수의 신뢰구간추정량에 대한 실제포함확률에 관한 연구

김대학¹

¹대구가톨릭대학교 인성교양부

접수 2010년 5월 20일, 수정 2010년 6월 5일, 게재확정 2010년 6월 10일

요약

본 연구는 이항분포의 성공의 확률에 대한 신뢰구간추정량들을 비교분석하고자 한다. 일반적으로 대표본의 경우에 적용되는 잘 알려진 신뢰구간추정량과 소표본의 경우에도 적용될 수 있는 정확신뢰구간, 그리고 포아송 분포를 이용하여 구한 신뢰구간추정량과 연속성의 수정을 고려한 추정량들을 소표본의 모의실험을 통하여 실제포함확률의 측면에서 비교하였다.

주요용어: 신뢰구간 추정, 실제포함확률, 이항모수, 이항분포.

1. 서론

오늘날의 복잡한 정보사회에서는 많은 자료가 생성되고 이들 중 상당 부분은 범주형으로 주어진다. 범주형 변수의 경우 개체들은 반응값이 성공 혹은 실패인 이항범주의 경우가 대부분을 차지한다. 특히 반응값이 이항인 경우 모비율 p 는 모집단에서 어떤 특정한 속성을 갖는 개체의 비율을 나타내는 모수라 할 수 있으며 이항모수로도 불리운다.

이제 크기가 n 인 랜덤표본에서 특정한 속성을 갖는 개수를 X 라 하고, 모비율의 추정량으로 표본비율 $\hat{p} = X/n$ 을 고려하자. 대부분의 통계학 교재에서 소개되고 있는 모비율의 신뢰구간추정은 대표본의 경우에 적용될 수 있는 Wald 신뢰구간만을 소개하고 있다. 이는 소 표본의 경우에 적용될 수 없는 한계점을 지니고 있으며 실제로 소 표본의 경우가 빈번하게 나타나고 있는 현실성을 감안할 때 실제 활용의 측면에서 사용가능한 신뢰구간을 소개하고 활용하는 것이 바람직해 보인다.

오래전부터 모비율 p 의 신뢰구간 구축과 관련된 연구가 진행되어 왔으며 1934년에 이르러 Clopper와 Pearson (1934)에 의해 이항모수에 대한 정확신뢰구간이 연구되었다. 최근 Agresti와 Coull (1998), Chen (1990) 그리고 Leemis와 Trivedi (1996) 등에 의해 여러 신뢰구간추정량들의 특징들이 계속 연구되고 있다. Leemis와 Trivedi (1996)는 모비율의 신뢰구간을 구축함에 있어 정규분포를 이용하는 방법과 포아송분포를 이용하는 방법을 비교하여 표본비율이 낮은 경우 포아송 신뢰구간을 활용하는 것이 좋을 있음을 보인다. 또한 Chen (1990)은 베이지안 추정을 이용하여 최적의 신뢰구간을 구축하는 방법을 제공하기도 하였다.

본 연구에서는 모비율 p 에 대한 신뢰구간추정량들을 살펴보고 모의실험을 통하여 실제포함확률 (actual coverage probability)의 측면에서 비교하였다. 이는 평균포함확률 (average coverage probability) 측면에서의 연구와 구분될 수 있다. 2절에서는 모비율 p 에 대한 여러 가지 신뢰구간 추정량을 살펴보고, 3절에서는 이들 신뢰구간추정량들을 예제를 통하여 실제로 계산하였고 4절에서는 실제포함확률 측면에서 모의실험을 통하여 비교하였다. 결론은 5절에 나타내었다.

¹ (712-702) 경상북도 경산시 하양읍 금락리 330, 대구가톨릭대학교 인성교양부, 교수.
E-mail: dhkim@cu.ac.kr

2. 모비율 p 의 신뢰구간 추정량

2.1. Clopper-Pearson 방법

확률변수 X 가 모비율이 p 인 이항분포를 따른다고 하자. 모비율 p 의 $100(1-\alpha)\%$ 신뢰구간을 구하는 방법으로서 Clopper와 Pearson (1934)은 다음과 같은 구조로 접근하였다. 확률변수의 실현값인 x 에 대해, 유의수준 $\alpha/2$ 에서 귀무가설 $H_0 : p = p_0$ 을 기각하지 않는 모든 p_0 를 계산함으로써 즉, 주어진 x 에 대해, 양측검정에서, 어떤 p_0 를 사용하여야 귀무가설을 채택할 수 있는지의 문제로 대체하여 신뢰구간 추정량을 구하였다. 이때 얻어지는 모든 p_0 중 최소값이 신뢰구간의 하한, 최대값이 신뢰구간의 상한이 된다. 즉, 최소값 p_L 과 최대값 p_U 는

$$P(X \geq x | p = p_L) = \sum_{k=x}^n \binom{n}{k} p_L^k (1-p_L)^{n-k} \leq \alpha/2$$

$$P(X \leq x | p = p_U) = \sum_{k=0}^x \binom{n}{k} p_U^k (1-p_U)^{n-k} \leq \alpha/2$$

로 계산된다. 이 신뢰구간을 Clopper-Pearson 신뢰구간, 혹은 정확신뢰구간이라고도 부른다.

위의 정확신뢰구간을 구하기 위해서는 상당히 많은 계산이 요구되지만 다행스럽게도 F 분포를 이용한 Blyth (1986)이나 Hald (1952)의 쉬운 계산방법이 존재한다. $x = 1, \dots, n-1$ 에 대하여 정확신뢰구간에 대한 닫힌 형태는 다음과 같다.

$$\left[1 + \frac{n-x+1}{xF_{2x, 2(n-x+1), 1-\alpha/2}} \right]^{-1} < p < \left[1 + \frac{n-x}{(x+1)F_{2(x+1), 2(n-x), \alpha/2}} \right]^{-1} \quad (2.1)$$

여기서, $F_{a,b,c}$ 는 자유도 a, b 를 따르는 F 분포의 $100(1-c)\%$ 분위점이다. $x=0$ 의 경우에는 하한을 0, $x=n$ 의 경우는 상한을 1로 두고 계산하면 된다.

2.2. 포아송분포를 이용하는 방법

모비율 p 의 또 다른 형태의 신뢰구간 추정량은 포아송분포로부터 유도될 수 있다. 포아송분포의 모수에 대한 정확신뢰구간으로부터 모비율 p 의 신뢰구간을 유도하는 방법은 다음과 같다. 포아송분포의 모수를 μ 라 할 때, μ 의 정확신뢰구간의 상한과 하한은 각각 다음을 만족하는 μ_U 와 μ_L 이다.

$$P(X \geq x | \mu = \mu_L) = \sum_{k=x}^n \frac{(e^{-\mu_L} \mu_L^k)}{k!} \leq \alpha/2$$

$$P(X \leq x | \mu = \mu_U) = \sum_{k=0}^x \frac{(e^{-\mu_U} \mu_U^k)}{k!} \leq \alpha/2$$

여기서, 포아송분포의 누적확률인 $P(X \leq x | \mu)$ 는 자유도가 $v = 2(1+x)$ 인 카이제곱분포 형태로 다음과 같이 표현된다.

$$\sum_{k=0}^x \frac{(e^{-\mu} \mu^k)}{k!} = P\{\chi_v^2 > 2\mu\}$$

위의 사실과, 이항분포의 포아송근사 (Poisson approximation)를 활용하여 $np \rightarrow \mu$ 로 놓으면, 다음과 같이 포아송분포의 정확확률로 모비율의 신뢰구간을 구할 수 있게된다. 즉

$$P(X \geq x | \mu = np_L) = \sum_{k=x}^{\infty} \frac{(e^{-np_L} (np_L)^k)}{k!} = 1 - \sum_{k=0}^{x-1} \frac{(e^{-np_L} (np_L)^k)}{k!} \leq \alpha/2$$

로 되고 포아송분포의 누적확률은 카이제곱분포를 이용하여

$$P\{\chi_{2x}^2 \geq 2np_L\} = 1 - \alpha/2$$

로 되어 신뢰구간의 하한 p_L 을 얻을 수 있으며, 비슷한 방법으로 상한 p_U 도 다음과 같이 계산된다.

$$p_L = \frac{1}{2n} \chi_{2x, \alpha/2}^2, p_U = \frac{1}{2n} \chi_{2(x+1), 1-\alpha/2}^2 \quad (2.2)$$

포아송모수의 신뢰구간의 성질에 대해서는 Leemis와 Trivedi (1996)에 의해 자세히 언급되었다. 특히 그들의 연구에서 표본의 크기와 표본비율이 주어질 때 정규근사를 사용하여야 하는지 포아송근사를 사용하여야 하는지의 경계점을 제시하기도 하였다.

2.3. 정규분포를 이용하는 방법

이제 표본비율 \hat{p} 의 분산을 $V(p) = \sqrt{p(1-p)/n}$ 로 놓으면, 확률변수 $(p - \hat{p})^2$ 는 정규근사 (normal approximation)에 의해 $(p - \hat{p})^2 = cV(\hat{p})$ 로 표현할 수 있다. 여기서, c 는 $\chi^2(1)$ 분포의 $100(1 - \alpha)\%$ 분위점이며, \hat{p} 는 p 에 대한 표현식이다. 이제, \hat{p} 에 적당한 값을 대입하여 위의 이차방정식을 풀면 두 값 $[p_L, p_U]$ 를 얻게된다. 잘 알려진 바대로, $\tilde{p} = \hat{p}$ 을 대입하여 방정식을 풀면, 모비율 p 에 대한 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n} \quad (2.3)$$

여기서, z_{α} 는 표준정규분포의 $100(1 - \alpha)\%$ 분위점이다. 이 신뢰구간을 Wald 신뢰구간이라고 하며, 대부분의 통계학 책에서 소개하고 있다.

2.4. 스코어 신뢰구간

스코어 (score) 신뢰구간은

$$(p - \hat{p})^2 = cV(\hat{p})$$

에서 $\tilde{p} = p$ 를 대입하여 이차방정식의 해로 얻어진다. 이는 주어진 귀무가설을 만족하는 신뢰영역 (acceptance region)을 구하는 것과 같은 문제인데, 이에 대한 신뢰구간은 다음과 같이 주어진다.

$$\frac{X + z_{\alpha/2}^2/2 \pm z_{\alpha/2} \sqrt{X - X^2/n + z_{\alpha/2}^2/4}}{n + z_{\alpha/2}^2} \quad (2.4)$$

위의 신뢰구간을 스코어 신뢰구간이라고 한다. Wald 신뢰구간은 분산 $V(p)$ 를 추정할 때 최대우도추정량 \hat{p} 을 이용하는 데 반해 스코어 신뢰구간은 Wilson (1927)이 지적한 바처럼, χ^2 분포의 일반화를 통해 p 를 이용한다는 점에서 두 신뢰구간은 서로 다른 모습을 하고 있다. Blyth와 Still (1983)은 스코어 신뢰구간의 성질에 대해서 자세히 논하였다.

2.5. 연속성 수정방법

2.3절의 정규분포를 이용하는 근사방법은 이산분포를 따르는 통계량을 연속확률분포인 정규분포로 근사함으로서 약간의 오차가 존재하리라고 예상할 수 있다. 그러므로 수정된 값을 대입시키는 연속성 수정 (continuity correction)으로 오차를 보정하는 방법을 생각할 수 있다. 즉, 이항분포를 따르는 확률변수 X 에 대해 $P(X = a)$ 의 정규근사 $P(a - 0.5 \leq Y \leq a + 0.5)$ 를 이용하는 방법을 고려할 수 있다. 여기서, Y 는 X 와 평균과 분산이 같은 정규분포이다. 이에 따라, 식 (2.3)에 연속성 수정을 하면 다음과 같은 신뢰구간이 주어진다.

$$\frac{X}{n} \pm \left\{ \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\left(\frac{X}{n}\right) \left(1 - \frac{x}{n}\right) + \frac{1}{2n}} \right\} \tag{2.5}$$

이외에도 이항모수의 신뢰구간을 추정하는 방법으로 arcsin 변환을 고려하는 신뢰구간, 붓스트랩 방법을 이용하는 신뢰구간 등을 고려할 수 있다.

3. 예 제

본 절에서는 2절에서 살펴본 여러 가지 신뢰구간 추정량들에 대한 계산결과를 살펴보고자 한다. 신뢰구간 추정량들은 2절에서 닫힌 형태 (closed form)로 나타나지만 그 계산과정이 한눈에 보일 정도로 쉽게 얻어지는 경우는 드물다.

표 3.1은 표본의수 n 이 작을 때 ($n = 5, 10, 20$) 주어진 값 x 에 대하여 95% 정확신뢰구간과 Wald 신뢰구간을 계산한 결과이다. 여기서 cp는 정확신뢰구간을, w는 Wald 신뢰구간을 의미한다.

표 3.1 추정 신뢰구간 ($\alpha = 0.05$)

n	x	cp	w
5	0	(0, .5218)	(0, 0)
	1	(0.0051, 0.7164)	(-.01506, 0.5506)
	2	(0.057, 0.8534)	(-.0294, 0.8294)
	3	(0.1466, 0.9473)	(0.1706, 1.0294)
	4	(0.2836, 0.9949)	(0.4494, 1.1506)
	5	(0.4782, 1)	(1, 1)
10	0	(0, 0.3085)	(0, 0)
	2	(0.0252, 0.5561)	(-0.0479, 0.4479)
	4	(0.1216, 0.7376)	(0.0964, 0.7036)
	6	(0.2624, 0.8784)	(0.2964, 0.9036)
	8	(0.4439, 0.9784)	(0.5521, 1.0479)
	10	(0.6915, 1)	(1, 1)
25	0	(0, 0.1372)	(0, 0)
	5	(0.0683, 0.4070)	(0.0432, 0.3568)
	10	(0.2113, 0.6133)	(0.2080, 0.5920)
	15	(0.3867, 0.7887)	(0.4080, 0.7920)
	20	(0.5930, 0.9317)	(0.6432, 0.9568)
	25	(0.8628, 1)	(1, 1)

표 3.1에서 알수 있듯이 Wald 신뢰구간은 표본의 수가 작은 경우 모비율의 신뢰구간의 하한이 음의 값도 가지는 경우도 있고 또한 신뢰구간의 상한이 1을 초과하는 경우도 생기게 됨을 알 수 있다. 한편 정확신뢰구간의 경우는 별 문제없이 신뢰구간을 추정하는 것처럼 보인다. 포아송분포를 이용하거나 연속성의 수정을 한 경우도 쉽게 구해질수 있다.

$n = 25$ 인 경우에 실현값 $x = 10$ 인 경우 우리는 모비율의 95% 신뢰구간을 (0.2113, 0.6133)로 추정하게 된다. 주어진 n 에 대하여 추정되는 신뢰구간의 길이는 실현값 x 가 가능한 값 전체영역의 가운데에 가까워질수록 신뢰구간의 길이는 커지는 경향을 나타내고 있음도 파악할 수 있다. 문제는 이 정확신뢰구간의 평균적 수행능력을 평가해 보는 것이다. 한 번의 경우에는 신뢰구간을 잘 추정하고 있는 것처럼 보이나 여러 번 반복하여 추정하여 볼 때 어떤 성질을 갖는지가 관건일 것이다. 이를 위하여 4절에서는 소표본 모의실험을 실시하여 보았다.

4. 소표본 모의실험

2절에서 소개한 모비율의 신뢰구간 추정량들의 효율을 비교하기 위하여 두가지 신뢰수준에서 또 소표본 ($n = 5, 10, 20, 30, 40$)의 경우에 모의실험을 실시하였다. 여러 가지 경우의 모비율 ($p = 0.2, 0.3, 0.5, 0.8$)에 대하여 각각 1000번의 반복을 통한 실제포함확률을 계산한 결과가 표 4.1에서 표 4.4까지 나타나 있다. 여기서 cp는 정확신뢰구간을, p는 포아송근사, w는 정규근사, s는 스코어방법 그리고 마지막으로 ad는 연속성 수정방법을 의미한다.

표 4.1 명목포함확률 ($\alpha = 0.05$, 반복수=1000)

모비율 방법 /n	p=0.2				p=0.3					
	5	10	20	30	40	5	10	20	30	40
cp	.994	.992	.970	.972	.976	.998	.991	.973	.977	.966
p	.646	.891	.970	.972	.985	.833	.967	.983	.986	.991
w	.646	.891	.919	.951	.950	.803	.834	.945	.958	.948
s	.925	.928	.930	.930	.935	.924	.925	.928	.930	.942
ad	.646	.891	.919	.951	.962	.803	.961	.945	.958	.966

표 4.2 명목포함확률 ($\alpha = 0.05$, 반복수=1000)

모비율 방법 /n	p=0.5				p=0.8					
	5	10	20	30	40	5	10	20	30	40
cp	.994	.992	.970	.972	.976	.992	.989	.990	.986	.971
p	.646	.891	.970	.972	.985	.999	.999	.999	.999	.999
w	.646	.891	.919	.951	.950	.662	.871	.930	.942	.950
s	.935	.938	.939	.945	.950	.934	.938	.940	.941	.945
ad	.646	.891	.919	.951	.962	.662	.871	.930	.942	.964

모의실험의 결과를 살펴보면 다음과 같이 요약된다. 정확신뢰구간은 모든 n 과 모든 p 의 경우에 명목 신뢰수준 0.95를 초과하고 있음을 발견할 수 있다. 즉 정확신뢰구간은 과대추정이 발생한다는 의미이다. 물론 n 이 커질수록 실제포함확률은 명목신뢰수준에 근접하고 있음도 알 수 있다. 포아송근사방법과 Wald 방법은 표본의 수가 작을 경우에는 큰 차이가 없으나 표본의 수가 커질수록 Wald 방법은 명목신뢰수준에 근접함을 발견하게 된다. 연속성의 수정을 한 방법도 wald 방법과 큰 차이없이 표본의 수가 커질수록 명목신뢰수준에 접근하고 있으나 표본의 수가 작은 경우에는 명목신뢰수준에 훨씬 떨어지는 포함확률을 보여주고 있다. 한편 스코어방법은 대부분의 경우 명목신뢰수준을 만족하고 있음을 알 수 있다.

신뢰수준이 90%인 경우는 표 4.3과 표 4.4에 나타나 있다. 이들의 결과는 신뢰수준이 95%인 경우와 거의 유사한 결과를 보여주고 있다.

한편, 모든 경우의 모비율 p 에 대하여 신뢰구간추정량들의 수행능력을 평가하기 위하여 0.01부터 0.01간격으로 0.99까지의 모비율에서 신뢰구간추정량들을 각각 1000번씩 반복하여 실제포함확률을 구

표 4.3 명목포함확률 ($\alpha = 0.1$, 반복수=1000)

모비율 방법 /n	p=0.2					p=0.3				
	5	10	20	30	40	5	10	20	30	40
cp	.994	.992	.970	.972	.976	.976	.918	.921	.939	.942
p	.646	.891	.970	.972	.985	.820	.956	.980	.984	.963
w	.646	.891	.919	.951	.907	.798	.789	.860	.882	.914
s	.855	.870	.890	.895	.899	.845	.860	.880	.890	.900
ad	.646	.891	.919	.951	.962	.798	.956	.855	.882	.914

표 4.4 명목포함확률 ($\alpha = 0.1$, 반복수=1000)

모비율 방법 /n	p=0.5					p=0.8				
	5	10	20	30	40	5	10	20	30	40
cp	.994	.992	.970	.972	.976	.987	.965	.957	.945	.912
p	.646	.891	.970	.972	.985	.991	.995	.998	.999	.999
w	.646	.891	.919	.951	.907	.588	.840	.919	.864	.860
s	.845	.869	.875	.885	.890	.845	.852	.863	.875	.890
ad	.646	.891	.919	.951	.962	.642	.871	.939	.945	.887

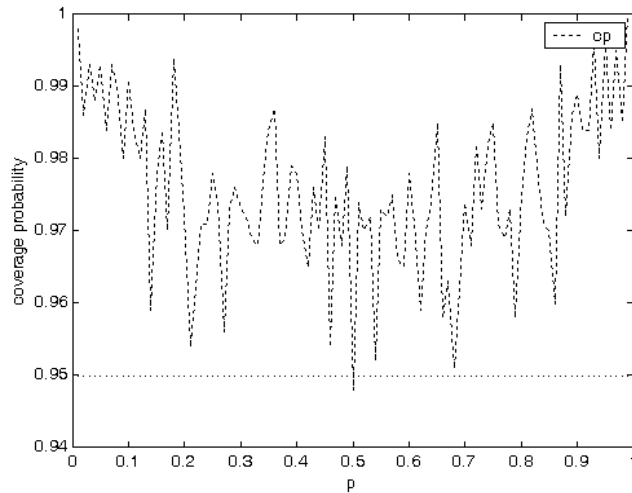


그림 4.1 정확신뢰구간 (cp)의 실제포함확률

하여 보았다. 그림 4.1은 정확신뢰구간의 실제포함확률을 나타내고 있다. 그림에서 알 수 있듯이 모든 모비율의 경우에 명목신뢰수준을 초과하고 있다. 이로부터 정확신뢰구간 추정량이 모비율을 과다추정하고 있음을 알 수 있다. 그림 4.2는 정확신뢰구간과 Wald 신뢰구간을 비교하고 있다. 이 그림으로부터 Wald 신뢰구간은 모비율이 아주 작거나 아주 클 때는 과소추정하고 있으며 모비율의 값이 커질수록 톱니 모양의 실제포함확률을 보여주고 있음을 알 수 있다.

표 4.5는 몇 가지 모비율 값에서 2절에서 논의한 여러 가지 신뢰구간 추정량들에 대한 실제포함확률을 나타내고 있다.

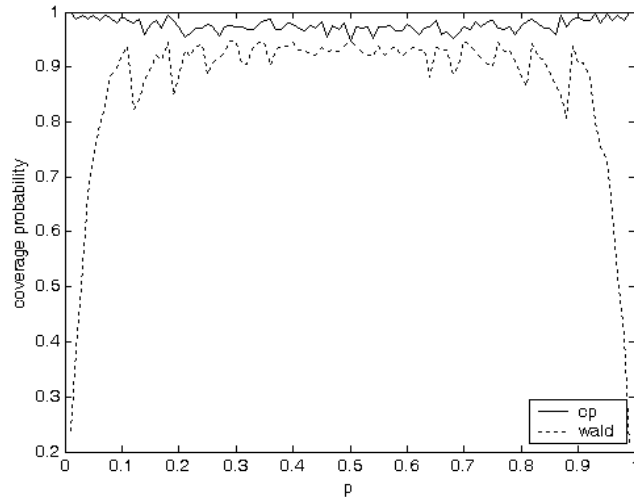


그림 4.2 정확신뢰구간과 wald 신뢰구간의 실제포함확률

표 4.5 명목포함확률 ($\alpha = 0.05, n = 25, \text{반복수}=1000$)

모비율 / 방법	cp	p	w	s	ad
0.01	.998	.237	.239	.940	.999
0.05	.993	.727	.733	.938	.733
0.10	.991	.922	.922	.941	.931
0.20	.972	.989	.882	.940	.969
0.30	.973	.983	.946	.942	.946
0.40	.978	.998	.945	.943	.964
0.50	.948	.999	.948	.945	.948
0.60	.965	.998	.920	.941	.965
0.70	.974	.998	.948	.944	.948
0.80	.975	.999	.881	.944	.957
0.90	.989	.999	.907	.942	.917
0.95	.996	.999	.731	.942	.731
0.99	.999	.999	.209	.941	.999

5. 결론

Agresti와 Coull (1998)은 Clopper와 Pearson (1934)의 정확신뢰구간보다 근사이론을 이용한 신뢰구간이 더 나을 수 있다는 것을 모의실험으로 보인바 있다. 그러나 이는 대표본의 경우에 적용할 수 있음을 기억하여야 한다. Clopper와 Pearson (1934)의 이 신뢰구간은 가장 표준적인 정확한 방법임에도 불구하고, 이항분포가 지니는 이산성에 의해 신뢰구간을 넓게 추정하는 경향을 모의실험을 통하여 발견할 수 있었다. 또한 모의실험에서 고려한 모든 경우에 스코어 신뢰구간이 실제 포함확률 측면에서 주어진 명목수준을 가장 잘 유지하고 있음을 발견하였다. 대표본의 경우에 적용할 수 있는 Wald 신뢰구간을 소표본의 경우에 적용할 때 발생하는 문제점도 쉽게 간과되어서는 안될 것으로 사료된다.

참고문헌

- Agresti, A. and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.
- Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78**, 108-116.
- Blyth, C. R. (1986). Approximate binomial confidence limits. *Journal of the American Statistical Association*, **81**, 843-855.
- Chen, H. (1990). The accuracy of approximate intervals for a binomial parameter. *Journal of the American Statistical Association*, **85**, 514-518.
- Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404-413.
- Hald, A. (1952). *Statistical theory with engineering applications*, John Wiley, New York.
- Leemis, L. M. and Trivedi, K. S. (1996). A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician*, **50**, 63-68.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.

On the actual coverage probability of binomial parameter

Daehak Kim¹

¹School of Liberal Arts, Catholic University of Daegu

Received 20 May 2010, revised 5 June 2010, accepted 10 June 2010

Abstract

In this paper, various methods for finding confidence intervals for the p of binomial parameter are reviewed. We compare the performance of several confidence interval estimates in terms of actual coverage probability by small sample Monte Carlo simulation.

Keywords: Actual coverage probability, binomial distribution, binomial parameter, confidence interval.

¹ (712-702) Professor, School of Liberal arts, Catholic University of Daegu, Kyungsan, Korea.
E-mail: dhkim@cu.ac.kr