Journal of the Korean Data & Information Science Society
2011, **22**(6), 1175–1182

한국데이터정보과학회지

실제포함확률을 이용한 초기하분포 모수의 근사신뢰구간 추정에 관한 모의실험 연구[†]

김대학[1]

[1]대구가톨릭대학교 수학과

접수 2011년 10월 31일, 수정 2011년 11월 18일, 게재확정 2011년 11월 22일

요 약

본 연구는 초기하분포의 모수, 즉 성공의 확률에 대한 신뢰구간추정에 대하여 설펴보았다. 초기하분포의 성공의 확률에 대한 신뢰구간은 일반적으로 잘 알려져 있지 않으나 그 응용성과 활용성의 측면에서 신뢰구간의 추정은 상당히 중요하다. 본 논문에서는 초기하분포의 성공의 확률에 대한 정확신뢰구간과 이항분포와 정규분포에 의한 근사신뢰구간을 소개하고 여러 가지 모집단의 크기와 표본 수에 대하여, 그리고 몇 가지 관찰값에 대한 정확신뢰구간과 근사신뢰구간을 계산하고 소 표본의 경우에 모의실험을 통하여 실제포함확률의 측면에서 살펴보았다.

주요용어: 실제포함확률, 이항근사, 정규근사, 정확신뢰구간, 초기하분포의 모수.

1. 서 론

희귀한 사건과 관련된 확률모형으로 자주 이용되는 포아송 분포와 함께 가장 기본적으로 사용되는 이산형 (discrete) 분포인 이항분포는 오늘날의 의학, 생물학적 응용에 있어서 사망자수 등과 관련된 자료의 분석에 있어서 많이 이용되고 있다. 또한 초기하분포는 모집단을 유한모집단으로 제한할 때 이항분포대신에 현실적으로 자주 이용되는 유용한 분포이다. 예를 들면 초기하분포의 응용은 제한된 수의 어린이들이 특정질병에 노출되었을 때 그 질병에 감염된 어린의 수에 대한 확률모형을 나타낼 때 사용될 수 있다. 또 다른 흥미로운 생물학적 예는 포획-방류 (capture and recapture)자료로부터 야생동물의 모집단의 크기를 추정할 때 사용될 수 있다. 또한 초기하 분포의 중요한 응용중의 하나는 다양한 의학보건조사 (biomedical survey)의 자료나 품질관리 (quality control)의 영역 등에서 이루어지고 있다. 특히 이항분포가 복원추출 (with replacement)의 경우 성공의 횟수 (number of success)의 확률분포 (probability distribution)에 해당된다면 초기하분포는 비복원추출 (without replacement)의 경우에 성공의 횟수에 관한 확률분포에 해당되는 이산형 분포로서 대부분의 학부 통계학 교재에서 소개되고 있는 잘 알려진 분포이다.

이제 p를 모집단에서의 특정속성을 지닐 확률로 정의하자. 즉 크기 N인 모집단에서 특정속성을 지닌 개체의 비율이라 하자. $\binom{N}{n}$개의 표본이 추출될 가능성이 동일하다고 전제하면 n개의 표본을 비복원 방법으로 추출할 경우 특정속성을 지닌 개체의 수인 확률변수 X가 관찰값 x가 될 확률은

$$P(X=x)=\frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}},\ x=\max(0,n-N(1-p)),\cdots,\min(n,Np) \tag{1.1}$$

[†] 이 연구는 2011년도 대구가톨릭대학교 교비연구비 지원에 의한 것임.
[1] (712-702) 경상북도 경산시 하양읍 하양로 13-13, 대구가톨릭대학교 수학과, 교수.
E-mail: dhkim@cu.ac.kr

로 됨은 잘 알려져 있다. 이제 크기가 n인 랜덤표본으로부터 특정한 속성의 확률 모비율 p의 추정을 고려하여 보자.

오래전부터 이항분포에서의 성공의 확률, 즉 이항 모비율 p의 신뢰구간 구축과 관련된 연구가 진행되어 왔으며 Clopper와 Pearson (1934)에 의해 이항모수에 대한 정확신뢰구간 (exact confidence interval)이 연구되었다. 그 후 Agresti와 Coull (1998), Chen (1990), Blyth와 Still (1983) 그리고 Leemis와 Trivedi (1996) 등에 의해 이항모수의 여러 신뢰구간 추정량들의 특징들이 계속 연구되고 있다. Leemis와 Trivedi (1996)는 모비율의 신뢰구간을 구축함에 있어 정규분포를 이용하는 방법과 포아송 (Poisson) 분포를 이용하는 방법을 비교하여 표본비율이 낮은 경우 포아송 신뢰구간을 활용하는 것이 좋음을 보인바 있다. 최근 Kim (2010a)은 이항모수의 신뢰구간 추정에 있어서 실제포함확률 (actual coverage probability)에 관하여 연구한 바 있다. 또한 Chen (1990)은 베이지안 추정 (Bayesian estimation)을 이용하여 최적의 신뢰구간 (confidence interval)을 구축하는 방법을 제공하기도 하였다. 그러나 앞서 언급한 바와 같이 실제로 의학이나 여러 응용분야에서 가장 활발히 이용되고 있는 초기하분포의 모수에 대한 신뢰구간추정은 학부과정에서 잘 언급되고 있지 않은 실정이다. 그 주된 이유는 학부생들이 직관적으로 이해하기 어려운 개념의 복잡성과 누적분포의 확률계산과 관련된 계산상의 어려움 때문으로 사료된다. 초기하 분포의 모수에 대한 신뢰구간추정은 Katz (1953)에 의해 시도된바 있으며 최근 컴퓨터의 발달과 더불어 계산능력의 실질적 개선을 이룬 프로그램을 이용하여 Sahai 와 Khurshid (1995)에 의해 구체화가 이루어졌고 Kim (2010b)은 초기하 분포의 모수에 대한 정확신뢰구간에 대하여 모의실험을 통하여 실제포함확률의 측면에서 연구한 바 있다.

본 연구에서는 초기하분포에서의 모비율 p에 대한 정확신뢰구간을 살펴보고 초기하분포의 특징상 이항분포를 이용한 근사를 이용하는 이항근사신뢰구간과 정규분포를 이용한 근사를 활용한 정규근사신뢰구간도 함께 살펴보았다. 정확신뢰구간과 근사신뢰구간의 특징을 다양한 모집단의 크기와 표본의 크기에 대하여 모의실험을 통하여 실제포함확률의 측면에서 비교하였다. 2절에서는 모비율 p에 대한 정확신뢰구간 추정량과 근사신뢰구간추정량들을 살펴보고, 3절에서는 이들 신뢰구간을 여러 가지 경우의 모집단의 크기와 표본크기, 그리고 몇가지 관찰값에 대하여 예를 통하여 실제로 계산하였고 4절에서는 실제포함확률 측면에서 소표본의 경우 모의실험의 결과를 나타내었다. 마지막으로 결론은 5절에 나타내었다.

2. 모비율 p의 정확신뢰구간과 근사신뢰구간

2.1. 정확신뢰구간

확률변수 X가 모비율이 p인 초기하분포를 따른다고 하자. 모비율 p의 $100(1-\alpha)\%$ 신뢰구간을 구하는 방법은 다음과 같은 가설검정 (hypothesis testing)의 구조로 설명가능하다. 다시 설명하면 주어진 확률변수 X의 관찰값 (observed value) x에 대해, 유의수준 $\alpha/2$에서 귀무가설 $H_0 : p = p_0$를 기각하지 않는 모든 p_0를 계산함으로서 신뢰구간을 추정할 수 있다. 즉, 주어진 x에 대해, 양측검정에서, 어떤 p_0를 사용하여야 귀무가설을 채택할 수 있는가 하는 문제로 대체하여 신뢰구간 추정량을 구하면 된다. 이때 얻어지는 모든 p_0중 최솟값 (minimum)이 정확신뢰구간의 하한, 최댓값 (maximum)이 정확신뢰

구간의 상한이 된다. 즉, 최솟값 p_L과 최댓값 p_U는

$$P(X \geq x|p = p_L) = \frac{\sum_{k=x}^{\min(n,Np_L)} \binom{Np_L}{k}\binom{N(1-p_L)}{n-k}}{\binom{N}{n}} = \frac{\alpha}{2}, \tag{2.1}$$

$$P(X \leq x|p = p_U) = \frac{\sum_{k=\max(0,n-N(1-p_U))}^{x} \binom{Np_U}{k}\binom{N(1-p_U)}{n-k}}{\binom{N}{n}} = \frac{\alpha}{2}. \tag{2.2}$$

를 만족하는 값으로 계산된다. 초기하 분포의 경우는 신뢰구간의 닫힌 형태 (closed form)가 없어 신뢰구간의 계산에 있어서는 실제로 모든 가능한 p에 대하여 식 (2.1)과 식 (2.2)를 만족하는 누적확률을 계산하여야만 우리가 원하는 정확신뢰구간을 얻을 수 있다. 물론 초기하분포가 가지는 특징 즉 관찰값 x보다 모집단에서의 특정속성을 지닌 개수가 더 커거나 같아야 하는 제한 때문에 이항분포나 포아송 분포의 경우와 같이 연속인 구간에서의 모든 확률에 대하여 계산할 수는 없는 상황을 전제하여야 한다. 이런 신뢰구간의 계산에는 엄청난 양의 계산이 요구되나 오늘날 발달한 컴퓨터 프로그램 (IMSL, 1994; Liberman and Owen, 1961; SAS, 1990)을 이용하면 신뢰구간을 비교적 쉽게 구할 수 있다. 식 (2.1)과 식 (2.2)를 만족하는 신뢰구간을 초기하 모수 p의 정확신뢰구간 (exact confidence interval)이라고도 부른다.

2.2. 이항근사신뢰구간

모비율 p가 작고 (일반적으로 $p < 0.1$) N이 상당히 큰 경우 ($N \geq 60$) 식 (1.1)은 이항분포에 의해 근사될 수 있다. 이항분포를 이용하여 초기하분포를 근사시킬 경우 식 (2.1)과 식 (2.2)는 다음의 식 (2.3)과 식 (2.4)로 근사된다.

$$P(X \geq x|p = p_L) = \sum_{k=x}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \frac{\alpha}{2}, \tag{2.3}$$

$$P(X \leq x|p = p_U) = \sum_{k=0}^{x} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \frac{\alpha}{2}. \tag{2.4}$$

식 (2.3)과 식 (2.4)를 p_L과 p_U에 관하여 풀면

$$p_L = \left[1 + \frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1}, \tag{2.5}$$

$$p_U = \left[1 + \frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1}. \tag{2.6}$$

를 얻게 된다. 이런 근사를 이항근사라고 표기하기로 하자. 여기서, $F_{a,b,c}$는 자유도 a, b를 따르는 F분포의 $100(1-c)\%$분위점이다. 여기서 주의할 점은 $x = 0$의 경우에는 하한을 0, $x = n$의 경우는 상한을 1로 두고 계산해야 한다는 점이다. 이항분포를 이용하여 근사시킬 경우에는 다행스럽게도 F분포를 이용한 Blyth (1986)이나 Hald (1952)의 쉬운 계산방법이 존재하여 정확신뢰구간의 닫힌 형태를 구할 수 있다.

2.3. 정규근사신뢰구간

초기하분포를 정규분포로 근사시키는 경우를 고려하자. 대략적으로 $np \geq 4$인 경우 확률함수 (1.1)은 평균이 np이고 분산이 $np(1-p)(N-n)/(N-1)$인 정규분포에 의해 근사될 수 있다. 이때 식 (2.1)과

식 (2.2)는 다음과 같이 표현가능하다.

$$P(X \geq x|p = p_{_L}) = P\left(Z \geq \frac{x - 0.5 - np_{_L}}{\sqrt{np_{_L}(1-p_{_L})\frac{N-n}{N-1}}}\right) = \frac{\alpha}{2}, \tag{2.7}$$

$$P(X \leq x|p = p_{_U}) = P\left(Z \leq \frac{x + 0.5 - np_{_U}}{\sqrt{np_{_U}(1-p_{_U})\frac{N-n}{N-1}}}\right) = \frac{\alpha}{2}. \tag{2.8}$$

식 (2.7)과 식 (2.8)을 $p_{_L}$과 $p_{_U}$에 관하여 풀면 다음의 식 (2.9)와 식 (2.10)을 얻게 된다.

$$p_{_L} = \frac{1}{2u}\left[u + (2x-n-1) - \sqrt{u^2 - \frac{2u}{n}\left\{(x-0.5)^2 + (n-x+0.5)^2\right\} + (2x-n-1)^2}\ \right], \tag{2.9}$$

$$p_{_U} = \frac{1}{2u}\left[u + (2x-n+1) + \sqrt{u^2 - \frac{2u}{n}\left\{(x+0.5)^2 + (n-x-0.5)^2\right\} + (2x-n+1)^2}\ \right]. \tag{2.10}$$

이때 $u = n + (1-n/N)z^2_{\alpha/2}$이고 $z_{\alpha/2}$는 표준정규분포의 상위 $100\cdot\alpha/2\%$분위점이다. 이런 정규근사를 정규근사1이라고 표현하기로 하자.

또 다른 형태의 정규근사도 가능하다. p가 너무 작지 않고 n이 충분히 클 때 표본비율 x/n을 평균이 p이고 분산이 $(x/n)(1-x/n)(N-n)/\{n(N-1)\}$인 정규분포를 따르는 변수로 근사시키는 방법이다. 이 방법은 간단하게 근사되는 장점이 있지만 정확성에 있어서는 약간의 손실이 발생하게 된다. 이 방법을 통하여 얻어지는 연속성의 수정이 가미된 모비율 p의 $100(1-\alpha)\%$신뢰구간은 다음의 식 (2.11)과 식 (2.12)로 얻어진다.

$$\frac{x}{n} - \left[z_{\alpha/2}\sqrt{\frac{N-n}{N(N-1)}\frac{x}{n}(1-\frac{x}{n}) + \frac{1}{2n}}\ \right], \tag{2.11}$$

$$\frac{x}{n} + \left[z_{\alpha/2}\sqrt{\frac{N-n}{N(N-1)}\frac{x}{n}(1-\frac{x}{n}) + \frac{1}{2n}}\ \right]. \tag{2.12}$$

이런 정규근사를 정규근사2라 표현하기로 하자.

3. 예 제

본 절에서는 2절에서 살펴본 초기하 모수 p의 정확신뢰구간과 근사신뢰구간에 대한 예제를 살펴보고자 한다. 신뢰구간 추정량은 2절에서 설명한 바와 같이 식 (2.1)과 식 (2.2)를 만족하는 구간이지만 그 계산과정이 한눈에 보일 정도로 쉽게 얻어지는 경우는 드물다.

표 3.1은 여러 가지 크기 ($N = 50, 100, 500$)의 유한모집단에 대하여 다양한 표본 수 n ($n = 10, 20, 50, 100, 450$)을 고려할 때 주어진 관찰값 x에 대한 95% 신뢰구간을 계산한 결과이다.

표 3.1의 모든 계산은 MATLAB 프로그램을 이용하여 구한 결과이다. 물론 FORTRAN 프로그램이나 SAS 프로그램 등을 이용하여 구할 수 있으나 그 결과는 큰 차이가 없음을 발견하였다. 표 3.1에서 모집단에서 표본을 추출하는 경우 표본의 크기는 각각의 N에서 n개의 표본을 얻도록 계획되었다. 표 3.1에서 제시된 모든 경우 구하여진 신뢰구간 추정 결과는 차이가 있음을 발견하게 된다. 다시 말하면 N이 50일 때 n이 10인 경우와 N이 100일 때 n이 20인 경우, 그리고 N이 500일 때 n이 100인 경우의 비율은 같으나 실현값 x에 따른 신뢰구간 추정결과는 서로 다르게 된다는 의미이다. 이는 주어진 실현값에 대하여 각 모집단에서의 성공의 확률이 서로 다른 이유 때문이다.

표 3.1 신뢰구간의 계산 ($\alpha = 0.05$)

			정확신뢰구간		정규근사1		이항근사	
N	n	x	p_L	p_U	p_L	p_U	p_L	p_U
50	10	3	0.14	0.62	0.091	0.619	0.067	0.653
		5	0.30	0.78	0.220	0.779	0.187	0.813
		7	0.48	0.92	0.381	0.909	0.348	0.933
	20	6	0.18	0.50	0.153	0.498	0.119	0.543
		10	0.36	0.68	0.317	0.683	0.272	0.728
		14	0.56	0.84	0.505	0.847	0.457	0.881
100	20	7	0.21	0.57	0.176	0.569	0.154	0.592
		12	0.43	0.79	0.384	0.786	0.361	0.809
		18	0.77	0.98	0.695	0.979	0.683	0.987
	50	11	0.16	0.32	0.142	0.321	0.115	0.359
		26	0.43	0.63	0.413	0.625	0.374	0.663
		41	0.08	0.89	0.722	0.891	0.685	0.914
500	100	30	0.23	0.39	0.222	0.391	0.212	0.399
		50	0.418	0.592	0.409	0.591	0.398	0.601
		70	0.620	0.781	0.609	0.777	0.601	0.787
	450	50	0.104	0.122	0.101	0.121	0.083	0.143
		200	0.212	0.236	0.428	0.460	0.397	0.492
		400	0.212	0.898	0.878	0.898	0.856	0.916

주어진 n에 대하여 추정되는 신뢰구간의 길이는 관찰값 x가 가질 수 있는 가능한 값 전체영역에 대하여 큰 변화가 없음을 알 수 있다. 여기서 우리의 관심은 이 정확신뢰구간과 근사신뢰구간들의 평균적 수행능력을 평가해 보는 것이다. 한 번 실험할 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 초기하 분포의 성공의 확률, 즉 모비율의 정확신뢰구간 추정량과 정규근사 신뢰구간 추정량, 그리고 이항근사 신뢰구간 추정량들의 효율을 비교하기 위하여 신뢰수준 α=0.05에서 모집단의 크기 N이 각각 50, 100 그리고 500일 때에 한하여 다양한 표본의 크기 n에 대하여 모의실험을 실시하였다. 이때 3가지 모비율 p (0.1, 0.5, 0.9)에 대하여 각각 1000번의 반복을 통한 실제포함확률을 계산한 결과가 표 4.1에서 표 4.3까지 나타나 있다.

표 4.1 명목포함확률

$N = 50$	정확방법			정규근사1			정규근사2			이항근사		
$n \setminus p$	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
5	.814	.824	.067	.989	.999	.998	.999	.974	.999	.999	.974	.386
10	.930	.959	.025	.995	.993	.991	.999	.932	.674	.943	.993	.665
15	.959	.960	.139	.959	.999	.969	.999	.934	.848	.944	.990	.846
20	.998	.970	.042	.990	.953	.995	.999	.953	.912	.930	.989	.912
25	.960	.940	.023	.999	.973	.999	.999	.973	.980	.941	.993	.980

모의실험의 결과를 살펴보면 다음과 같이 요약된다. N이 50인 경우와 N이 100인 경우에는 정확방법에 의한 신뢰구간은 p가 0.9와 같이 큰 경우 그 수행능력이 아주 저조함을 알 수 있다. 이 경우 p가 0.1인 경우의 결과와는 아주 대조적으로 나타난다. 물론 이항근사를 이용한 경우에도 그 수행능력은 명

표 4.2 명목포함확률

$N = 100$	정확방법			정규근사1			정규근사2			이항근사		
$n \setminus p$	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
10	.907	.944	.259	.989	.984	.991	.999	.912	.665	.989	.984	.658
20	.973	.931	.633	.973	.964	.969	.999	.964	.909	.996	.964	.905
30	.958	.973	.614	.992	.955	.993	.999	.955	.982	.999	.994	.982
40	.962	.967	.644	.991	.978	.991	.964	.978	.964	.999	.991	.996
50	.942	.952	.618	.983	.979	.985	.958	.979	.961	.999	.999	.999

표 4.3 명목포함확률

$N = 500$	정확방법			정규근사1			정규근사2			이항근사		
$n \setminus p$	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
50	.954	.958	.963	.984	.978	.981	.959	.941	.977	.984	.978	.981
100	.947	.957	.945	.947	.965	.966	.943	.965	.951	.966	.981	.977
150	.951	.953	.960	.967	.965	.975	.954	.965	.969	.991	.988	.996
200	.965	.943	.965	.951	.953	.953	.940	.953	.947	.995	.989	.993
250	.954	.942	.967	.970	.951	.980	.959	.951	.967	.997	.997	.999

목확률에 못 미치고 있음도 알수 있었다. 모비율이 작고 N이 큰 경우 이항근사에 의한 방법은 정확신뢰구간의 경우보다 더욱 더 명목확률을 과도하게 초과하고 있으며 정규근사의 경우 대부분의 경우 주어진 명목 신뢰수준을 초과하고 있음을 알 수 있었고 N이 500과 같이 아주 큰 경우에는 정확방법이 모든 고려된 모비율의 경우에 명목신뢰수준에 아주 가깝게 접근함을 확인 할 수 있었다. 정규근사2의 경우는 정규근사1의 방법과 대표본의 경우 큰 차이가 없음을 발견할 수 있었지만 소표본의 경우에 다소 명목신뢰수준에 모자라게 됨을 알 수 있었다. 정확신뢰구간의 경우 모비율이 크고 ($p > 0.7$) n이 작은 경우 명목포함확률은 명목신뢰수준 95%를 하향하고 있음을 발견할 수 있다. 즉 정확신뢰구간은 과소추정이 발생한다는 의미이다. 물론 n이 커질수록 실제포함확률은 명목신뢰수준에 근접하고 있음도 알 수 있다. 또한 모비율이 작은 경우 ($p > 0.1$)에도 마찬가지로 명목신뢰수준을 하향하고 있음을 발견할 수 있다.

본 논문에서 결과를 제시하지는 않았지만 신뢰수준이 90%인 경우의 결과는 신뢰수준이 95%인 경우와 거의 유사하게 나타남을 모의실험을 통하여 확인할 수 있었다. 모비율이 크고 표본의 크기가 작은 경우와 모비율이 작은 경우 명목포함확률이 명목신뢰수준을 하향함을 발견할 수 있다. 물론 이런 현상은 신뢰수준이 90%일 때 에도 발생하였다.

5. 결 론

본 논문에서는 초기하 분포의 모수에 대한 신뢰구간추정량들을 비교, 연구하였다. 정확신뢰구간과 이항분포와 정규분포를 이용한 근사신뢰구간 추정량의 성질을 실제포함확률의 측면에서 모의실험을 통하여 연구하였다. 이항분포의 경우 Agresti와 Coull (1998)은 Clopper와 Pearson (1934)의 정확신뢰구간보다 근사이론을 이용한 신뢰구간이 더 나을 수 있다는 것을 모의실험으로 보인바 있다. 본 연구의 결과, 초기하분포의 경우에도 초기하분포가 지니는 이산성에 의해 정확신뢰구간 추정량은 신뢰구간을 넓게 추정하는 경향을 모의실험을 통하여 발견할 수 있었으며 근사신뢰구간을 이용한 신뢰구간의 추정이 정확신뢰구간보다 대표본이 전제된다면 더 나을 수 있음을 알 수 있었다..

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. (1986). Approximate binomial confidence limits. *Journal of the American Statistical Association*, **81**, 843-855.

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

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.

IMSL. (1994). *International mathematical and statistical libraries* (*FORTRAN subroutines for evaluating special functions*), Version 3, Visual Numerics, Houston, Texas.

Katz, L. (1953). Confidence intervals for the number showing a certain characteristic in a population when sampling is without replacement. *Journal of American Statistical Association*, **48**, 256-261.

Kim, D. (2010a). On the actual coverage probability of binomial parameter. *Journal of the Korean Data & Information Science Society*, **21**, 737-745.

Kim, D. (2010b). On the actual coverage probability of hypergeometric parameter. *Journal of the Korean Data & Information Science Society*, **21**, 1109-1115.

Leemis, L. M. and Trivedi, K. S. (1996). A comparison of approximate interval estimators for the bernoulli parameter. *The American Statistician*, **50**, 63-68.

Liberman, G. J. and Owen, D. B. (1961). *Tables of the hypergeometric probability distributions,* Stanford University Press, Stanford.

Sahai, H. and Khurshid, A. (1995). A note on confidence intervals for the hypergeometric parameter in analyzing biomedical data. *Computational Biological Medicine*, **25**, 35-38.

SAS. (1990). *SAS language : Reference,* Version 6, First Edition, SAS Institute, Cary, North Carolina.

Journal of the Korean Data & Information Science Society
2011, **22**(6), 1175–1182

한국데이터정보과학회지

A simulation study for the approximate confidence intervals of hypergeometric parameter by using actual coverage probability[†]

Daehak Kim[1]

[1] Department of mathematic, Catholic University of Daegu

Received 31 October 2011, revised 18 November 2011, accepted 22 November 2011

Abstract

In this paper, properties of exact confidence interval and some approximate confidence intervals of hyper-geometric parameter, that is the probability of success p in the population is discussed. Usually, binomial distribution is a well known discrete distribution with abundant usage. Hypergeometric distribution frequently replaces a binomial distribution when it is desirable to make allowance for the finiteness of the population size. For example, an application of the hypergeometric distribution arises in describing a probability model for the number of children attacked by an infectious disease, when a fixed number of them are exposed to it. Exact confidence interval estimation of hypergeometric parameter is reviewed. We consider the approximation of hypergeometirc distribution to the binomial and normal distribution respectively. Approximate confidence intervals based on these approximation are also adequately discussed. The performance of exact confidence interval estimates and approximate confidence intervals of hypergeometric parameter is compared in terms of actual coverage probability by small sample Monte Carlo simulation.

Keywords: Actual coverage probability, binomial approximation, exact confidence interval, hyper-geometric parameter, normal approximation.

[†] This work was supported by research grants from the Catholic University of Daegu in 2011.
[1] Professor, Department of Mathematics, Catholic University of Daegu, Kyungsan 712-702, Korea. E-mail: dhkim@cu.ac.kr