Journal of the Korean Data & Information Science Society
2010, **21**(6), 1109–1115

한국데이터정보과학회지

초기하분포의 모수에 대한 신뢰구간추정

김대학[1]

[1]대구가톨릭대학교, 인성교양부

접수 2010년 9월 7일, 수정 2010년 9월 7일, 게재확정 2010년 10월 25일

요 약

본 연구는 질병자료나 사망자수 등과 관련된 자료의 분석에서 가장 많이 사용되는 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 정확신뢰구간을 소개하고 여러 가지 모집단의 크기와 표본수에 대하여, 그리고 몇가지 실현값에 대한 신뢰구간을 유도하고 소표본의 경우에 모의실험을 통하여 실제 포함확률의 측면에서 살펴보았다.

주요용어: 실제포함확률, 정확신뢰구간, 초기하분포, 초기하분포의 모수.

1. 서 론

오늘날의 의학, 생물학적 응용에 있어서 이항분포는 질병자료나 사망자수 등과 관련된 자료의 분석에 있어서, 희귀한 사건과 관련된 확률모형으로 자주 이용되는 포아송 분포와 함께 가장 기본적으로 사용되는 이산형 (discrete) 분포이다. 한편 초기하분포는 모집단을 유한모집단으로 제한할 때 이항분포대신에 현실적으로 자주 이용되는 유용한 분포이다. 예를 들면 초기하분포의 응용은 제한된 수의 어린이들이 특정질병에 노출 되었을때 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 방법 으로 추출할 경우 특정속성을 지닌 개체의 수 X가 x가 될 확률은

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

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

로 됨은 잘 알려져 있다. 이제 크기가 n인 랜덤표본으로 부터 특정한 속성의 확률 모 비율 p의 추정을 고려하여 보자.

오래전부터 이항분포에서의 성공의 확률, 즉 이항 모 비율 p의 신뢰구간 구축과 관련된 연구가 진행되어 왔으며 Clopper와 Pearson (1934)에 의해 이항모수에 대한 정확신뢰구간 (exact confidence interval)이 연구되었다. 최근 Agresti와 Coull (1998), Chen (1990), Blyth 와 Still (1983) 그리고 Leemis와 Trivedi (1996) 등에 의해 여러 신뢰구간추정량들의 특징들이 계속 연구되고 있다. Leemis와 Trivedi (1996)는 모 비율의 신뢰구간을 구축함에 있어 정규분포를 이용하는 방법과 포아송 (Poisson) 분포를 이용하는 방법을 비교하여 표본비율이 낮은 경우 포아송 신뢰구간을 활용하는 것이 좋음을 보인바 있다. Kim (2010)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 초기하 분포의 모수에 대한 신뢰구간추정은 잘 언급되고 있지 않은 실정이다. 그 주된 이유는 학부생들이 직관적으로 이해하기 어려운 개념의 복잡성과 누적분포의 확률계산과 관련된 계산상의 어려움 때문으로 사료된다.

본 연구에서는 초기하분포에서의 모 비율 p에 대한 정확신뢰구간추정을 살펴보고 다양한 모집단의 크기와 표본의 크기에 대하여 모의실험을 통하여 실제포함확률 의 측면에서 비교하였다. 초기하 분포의 신뢰구간추정에 관해서는 Katz (1956)에 의해 시도된바 있다. 2절에서는 모 비율 p에 대한 정확신뢰구간 추정량을 살펴보고, 3절에서는 이들 정확신뢰구간을 여러 가지 경우의 모집단의 크기와 표본크기, 그리고 몇가지 실현값에 대하여 예를 통하여 실제로 계산하였고 4절에서는 실제포함확률 측면에서 소표본의 경우 모의실험의 결과를 나타내었다. 마지막으로 결론은 5절에 나타내었다.

2. 모비율 p의 정확신뢰구간추정

확률변수 X가 모 비율이 p인 초기하분포를 따른다고 하자. 모 비율 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 최소값 (minimum)이 정확신뢰구간의 하한, 최대값 (maximum)이 정확신뢰구간의 상한이 된다. 즉, 최소값 p_L과 최대값 p_U는

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

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

를 만족하는 값으로 계산된다. 이 신뢰구간을 구하기 위해서는 가능한 모든 확률에 대하여 초기하분포의 누적분포함수를 계산하여 우리 가 위의 조건을 만족하는 모 비율을 추정하면 된다. 물론 초기하분포

가 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; Liberman과 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.

모집단의 분포가 이항분포일 때 이항모수의 신뢰구간추정문제는 다행스럽게도 F분포를 이용한 Blyth (1986)이나 Hald (1952)의 쉬운 계산방법이 존재하여 정확신뢰구간의 닫힌 형태를

$$\left[1+\frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}\right]^{-1} < p < \left[1+\frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}\right]^{-1} \tag{2.3}$$

단 $F_{a,b,c}$는 자유도 a, b를 따르는 F분포의 $100(1-c)\%$분위점

에 의해 구할 수 있으나 초기하 분포의 경우는 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.

3. 예 제

본 절에서는 2절에서 살펴본 초기하 모수 p의 정확신뢰구간 추정량에 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 = 5, 10, 20)을 고려할 때 주어진 관찰값 x들에 대한 95% 정확신뢰구간을 계산한 결과이다.

표 3.1 정확신뢰구간 ($\alpha = 0.05$)

N	n	x	p_L	p_U	구간길이
50	10	3	0.14	0.62	0.48
		5	0.30	0.78	0.48
		7	0.48	0.92	0.44
	20	8	0.08	0.60	0.52
		10	0.36	0.70	0.34
		15	0.50	0.82	0.32
100	20	3	0.07	0.36	0.29
		5	0.13	0.47	0.34
		7	0.21	0.57	0.36
	40	8	0.13	0.32	0.19
		10	0.17	0.38	0.21
		15	0.28	0.51	0.23
500	100	3	0.014	0.080	0.066
		5	0.026	0.106	0.080
		7	0.038	0.132	0.094
	200	8	0.026	0.070	0.044
		10	0.032	0.082	0.050
		15	0.054	0.110	0.056

표 3.1의 모든 계산은 MATLAB 프로그램을 이용하여 구한 결과이다. 물론 FORTRAN 프로그램이나 SAS 프로그램등을 이용하여 구할수 있으나 그 결과는 큰 차이가 없음을 발견하였다. 표 3.1에서 모

집단에서 표본을 추출하는 경우 표본의 크기는 각각의 N에서 같은 구조를 가지도록 계획되었다. 그러나 각각의 경우 구하여진 그 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 x가 3일때의 계산된 신뢰구간은 서로 다르게 된다는 의미이다. 이는 주어진 실현값에 대하여 각 모집단에서의 성공의 확률이 서로 다른 이유 때문이다.

주어진 n에 대하여 추정되는 신뢰구간의 길이는 실현값 x가 가능한 값 전체영역의 서 큰 변화가 없음을 알수 있다. 여기서 우리의 관심은 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 초기하 분포의 성공의 확률, 즉 모 비율의 정확신뢰구간 추정량의 효율을 비교하기 위하여 두 가지 신뢰수준 ($\alpha = 0.05,\ \alpha = 0.1$)에서 모집단의 크기 N가 각각 50, 100 그리고 500일 때에 한하여 다양한 표본의 크기 n ($n = 5,\ 10,\ 20,\ 30,\ 40$)에 대하여 모의실험을 실시하였다. 이때 모비율 9 가지 ($p = 0.1,\ 0.2,\ \cdots,\ 0.8,\ 0.9$)에 대하여 각각 1000번의 반복을 통한 실제포함확률을 계산한 결과가 표 4.1에서 표 4.5까지 나타나 있다.

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

$N = 50$	p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	.814	.915	.964	.915	.974	.919	.820	.674	.386
10	.930	905	.977	.961	.961	.963	.843	.906	.665
15	.978	.994	.954	.946	.960	.938	.899	.876	.846
20	.918	.979	.928	.963	.970	.913	.930	.972	.907
25	.960	.973	.944	.977	.955	.971	.941	.962	.980

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

$N = 100$	p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	.902	.975	.962	.955	.946	.948	.880	.885	.649
20	.978	934	.938	.954	.962	.948	.976	.923	.901
30	.991	.986	.917	.967	.958	.915	.939	.982	.966
40	.968	.969	.961	.946	.951	.960	.953	.952	.958
50	.947	.961	.940	.947	.950	.941	.949	.954	.945

표 4.3 명목포함확률 ($\alpha = 0.05$, 반복수=1000)

$N = 500$	p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	.883	.959	.946	.947	.932	.935	.960	.870	.630
100	.950	.968	.943	.934	.941	.919	.957	.916	.730
150	.934	.950	.957	.926	.946	.928	.958	.938	.819
200	.963	.939	.961	.952	.936	.947	.944	.907	.912
250	.952	.966	.941	.940	.954	.929	.933	.941	.632

모의실험의 결과를 살펴보면 다음과 같이 요약된다. 모 비율이 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 모 비율이 작은 경우 ($p > 0.1$)에도 마찬가지로 명목신뢰수준을 하향하고 있음을 발견할 수 있다.

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

$N = 100$	p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	.892	.890	.875	.820	.794	.800	.834	.861	.639
20	.873	.949	.892	.861	.839	.858	.885	.935	.873
30	.957	.899	.914	.906	.912	.912	.927	.893	.858
40	.963	.891	.874	.900	.899	.888	.923	.859	.959
50	.970	.939	.907	.893	.878	.889	.924	.951	.947

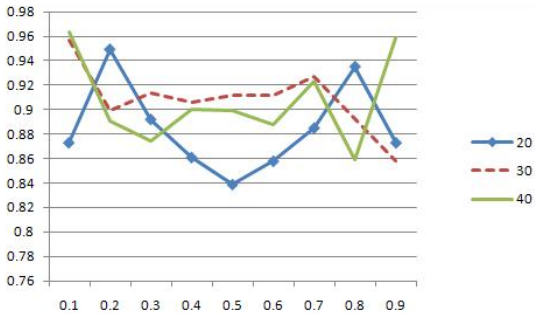

그림 4.1 모 비율의 변화에 따른 실제포함확률 ($N = 100$)

신뢰수준이 90%인 경우는 표 4.4과 표 4.5에 나타나 있다. 이들의 결과는 신뢰수준이 95%인 경우와 거의 유사한 결과를 보여주고 있다. 모 비율이 크고 표본의 크기가 작은 경우와 모비율이 작은 경우 명목포함확률이 명목신뢰수준을 하향함을 발견할 수 있다.이 90%일때에도 발생함을 알수 있다. 한편 그림 4.1은 모집단의 크기 N이 100일 때 다양한 표본의 크기에 대하여 모비율의 변화에 대한 실제포함확률을 나타내고 있다. 이 그림으로부터 알수 있듯이 모비율이 작거나 크게되는 양쪽 끝부분에서의 실제포함확률이 표본의 크기에 따라서 급격하게 변함을 확인 할 수 있다. 그림 4.2는 모집단의 크기 N이 500일때의 결과이다.

표 4.5 명목포함확률 ($\alpha = 0.1$, 반복수=1000)

$N = 500$	p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
50	.916	.903	.883	.860	.927	.848	.881	.879	.883
100	.891	.887	.902	.891	.880	.889	.904	.908	.883
150	.891	.906	.904	.875	.927	.872	.908	.910	.912
200	.937	.888	.885	.902	.886	.910	.899	.876	.935
250	.918	.931	.901	.879	.881	.896	.891	.925	.936

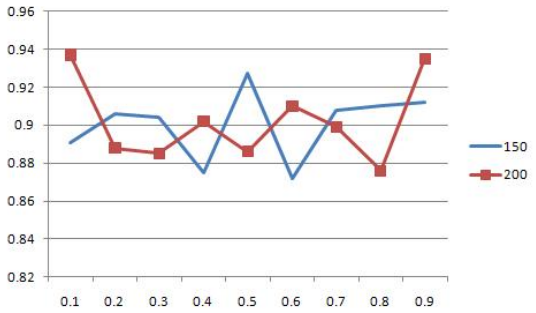

그림 **4.2** 모 비율의 변화에 따른 실제포함확률 ($N = 500$)

5. 결 론

Agresti와 Coull (1998)은 Clopper와 Pearson (1934)의 정확신뢰구간보다 근사이론을 이용한 신뢰구간이 더 나을 수 있다는 것을 모의실험으로 보인바 있다. 그러나 이는 대표본의 경우에 적용할수 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. (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. (2010). On the actual coverage probability of binomial parameter. *Journal of the Korean Data & Information Science Society*, **21**, 737-745.

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.

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

Journal of the Korean Data & Information Science Society
2010, **21**(6), 1109–1115

한국데이터정보과학회지

On the actual coverage probability of hypergeometric parameter

Daehak Kim[1]

[1]School of Liberal Arts, Catholic University of Daegu

Received 7 September 2010, revised 7 September 2010, accepted 25 October 2010

Abstract

In this paper, exact confidence interval 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 performance of exact confidence interval estimates of hypergeometric parameter in terms of actual coverage probability by small sample Monte Carlo simulation.

Keywords: Actual coverage probability, confidence interval, hyper-geometric distribution, hyper-geometric parameter.

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