시스템의 확률 값 시험을 위한 신뢰구간 비교 분석

황익순

Comparison of confidence intervals for testing probabilities of a system

Ik-soon Hwang

요 약

확률적 특성을 가지는 시스템의 시험을 위해서는 시험 입력을 일정 횟수만큼 반복하여 제공하고 관찰된 데이터를 기반으로 판정이 내려져야 한다. 구간 추정 기법을 이용하여 관찰된 데이터로부터 확률 값이 올바른지 여부를 판단할 수 있으며, 이 때 적절한 신뢰구간의 선택은 시험의 품질을 결정하는 중요한 요인이 된다. 본 논문에서는 다양한 크기의 표본에 대해 대표적인 구간 추정 기법인 Wald 신뢰구간과 Agresti-Coull 신뢰구간을 비교 분석한다. 각 신뢰구간이 확률 값 시험에 사용되었을 경우 올바른 구현 제품이 시험을 통과할 확률과 잘못된 구현제품이 시험을 통과하지 못할 확률을 기반으로 비교 분석을 수행하며, 확률 값이 올바른지를 판단하기 위한 양측검정뿐만 아니라 확률 값이 기준 확률 이상인지 여부를 판단하기 위한 단측검정을 사용하는 경우에 대해서도 비교 분석을 수행한다. 비교 분석 결과 양측검정의 경우 Agresti-Coull 신뢰구간을 사용할 것을 추천하며, 단측검정의 경우 큰 크기의 표본에 대해서는 Agresti-Coull 신뢰구간을, 적은 크기의 표본에 대해서는 Wald 신뢰구간 또는 Agresti-Coull 신뢰구간을 선택적으로 사용할 것을 추천한다.

ABSTRACT

When testing systems that incorporate probabilistic behavior, it is necessary to apply test inputs a number of times in order to give a test verdict. Interval estimation can be used to assert the correctness of probabilities where the selection of confidence interval is one of the important issues for quality of testing. The Wald interval has been widely accepted for interval estimation. In this paper, we compare the Wald interval and the Agresti-Coull interval for various sizes of samples. The comparison is carried out based on the test pass probability of correct implementations and the test fail probability of incorrect implementations when these confidence intervals are used for probability testing. We consider two-sided confidence intervals to check if the probability is close to a given value. Also one-sided confidence intervals are considered in the comparison in order to check if the probability is not less than a given value. When testing probabilities using two-sided confidence intervals, we recommend the Agresti-Coull interval. For one-sided confidence intervals, the Agresti-Coull interval is recommended when the size of samples is large while either one of two confidence intervals can be used for small size samples.

키워드

확률 시험, 구간 추정, 신뢰 구간, Wald 신뢰구간, Agresti-Coull 신뢰구간

Keywords

Probability testing, Interval estimation, Confidence interval, Wald Interval, Agresti-Coull interval

Software-Networks Department, TELECOM SudParis, France (Iksoon.Hwang@it-sudparis.eu)

접수일자 : 2010. 08. 31 심사(수정)일자 : 2010. 09. 30 게재확정일자 : 2010. 10. 15

Ⅰ. 서 론

최근 들어 공평성(fairness) 및 강건성(robustness)을 제공하기 위해 비결정성(nondeterminism) 또는 확률적 동작을 가지는 통신 프로토콜이 증가하고 있다. 예를 들어 [1]에서 제안된 프로토콜에서는 서비스의 공평성을 위한 부인-봉쇄(non-repudiation) 기능을 제공하기 위해 암호화 된 메시지를 복호화(decryption)하기 위한 키(key)가 정해진 확률 값에 따라 송신된다. 이러한 비결정성 및 확률적 동작은 시스템을 추상화하였을 경우에도 발생할 수 있다. 예를 들어 음성 통화 서비스를 제공하는 전체 시스템은 통화 시도라는 입력에 대해 일정한 확률로 통화 성공 또는 통화 실패라는 출력을 내는 고도로 추상화된 모델로 표현할 수 있으며, 이 때 각 확률 값은 통화 성공률과 통화 실패율이 된다. 이러한 확률적 특성을 가지는 구현 제품 및 시스템의 시험에 있어서 시험 대상의 정확한 확률 값이 알려져 있지 않을 경우 이를 시험하기 위해서는 시험 대상으로부터 관찰된 데이터를 기반으로 판정이 이루어져야 한다. 시험 대상의 내부를 볼 수 없는 블랙 박스 시험의 경우가 대표적인 예라 할 수 있으며, 성능 시험과 같이 시스템의 통계적 특성을 시험하는 경우에도 제한된 크기의 측정 데이터를 기반으로 판정이 내려져야 한다.

구간 추정(interval estimation) 기법은 특정한 분포 함수(distribution function)를 가지는 모집단에 대해 분포 함수의 파라미터 값을 모집단으로부터 관찰된 데이터, 즉 표본(sample)을 기반으로 하여 일정 수준의 신뢰수준(confidence level)를 가지고 구간으로 추정하는 기법이다. 구간 추정 기법 중 일반적으로 널리 알려진 방법은 Central Limit Theorem(CLT)을 기반으로 한 Wald 신뢰구간(confidence interval)이다 [2]. 비록 Wald 신뢰구간이 널리 사용되어 왔지만 Wald 신뢰구간은 추정하고자 하는 확률 값이 0 또는 1에 근접할 경우 낮은 커버리지 확률 (coverage probability: 해당 신뢰 구간이 실제 확률 값을 포함하고 있을 확률)을 제공하는 등의 문제점이 있는 것으로 알려져 있다 [3, 4]. 이러한 문제점을 해결하기 위해 다양한 신뢰구간이 제안되었으며, [3]과 [4]에서는 각각 양측검정과 단측검정을 사용하는 경우 다양한 신뢰구간에 대한 비교 분석을 수행하여 양측검정 사용 시에는 Agresti-Coull 신뢰구간 [5]을 포함한 몇 개의 신뢰구간을 사용할 것을 추천하였고, 단측검정 사용시에는 Jeffreys 신뢰구간 [3]을 사용할 것을 추천하였다. 하지만 최적의 신뢰구간에 대해서는 의견의 차이를 보이고 있으며, 신뢰구간의 비교를 위한 적절한 기준의 설정 및 설정된 기준의 적합성에 대한 논의가 계속되고 있다 [6].

[7]에서는 구간 추정 기법을 이용하여 PFSM (Probabilistic Finite-State Machine)을 시험하는 방안이 제안되었다. 제안된 방안에서는 PFSM에 시험 입력을 일정 횟수만큼 반복하여 제공하고 관찰된 데이터를 기반으로 PFSM의 각 천이(transition)의 확률 값이 올바르게 구현되어 있는지 여부를 Agresti-Coull 신뢰구간을 이용하여 판단하였다. 시험의 품질(quality)을 판단하기 위해 올바른 구현 제품이 시험을 통과할 확률(P_{EP})과 잘못된 구현제품이 시험을 통과하지 못할 확률(P_{NEF})을 고려하였으며, P_{EP}와 P_{NEF}를 보장하기 위한 시험 입력 반복 횟수를 결정하는 방안을 제시하였다. [8]에서는 [7]에서 제안한 기법을 이용하여 이동통신 시스템의 통화 성공률이 성능 요구사항을 만족하는지 여부를 시험하는 방안을 제안하였다. [7]에서는 시험 대상의 확률 값이 기준 확률 값과 유사한지 여부를 판단해야 하므로 양측검정을 사용한 반면, [8]에서는 시험 대상의 확률 값이 기준 확률 값 이상인지를 판단해야 하므로 단측검정을 사용한 차이점이 있다.

구간 추정 기법을 이용하여 확률 값을 시험하는 경우 시험 특성에 맞는 적절한 신뢰구간의 선택이 매우 중요하다. [3]과 [4]에서 수행한 신뢰구간 비교 분석은 기본적으로 커버리지 확률과 신뢰구간의 길이를 기반으로 하고 있다. Wald 신뢰구간의 경우 낮은 커버리지 확률 값을 가지는 몇 가지 확률 값과 표본 크기의 조합에 대한 실험 결과를 바탕으로 사용하기에 적합하지 않음을 보였으며, 대부분의 경우 200 이하의 표본 크기에 대해 신뢰구간의 커버리지 확률 등을 비교 분석하였다. 신뢰구간을 이용하여 확률 값을 시험할 경우 표본의 크기는 곧 시험 입력 반복 횟수가 되며, [7]에서 보인 것과 같이 높은 품질의 시험 결과를 얻기 위해서는 시험 입력 반복 횟수는 200보다 큰 값을 가지는 경우가 많다. 특히 [8]에서와 같이 시스템의 성능 요구사항을 시험하는 경우에는 10,000 이상의 크

기를 가지는 표본이 사용되는 경우도 자주 발생한다. [3]과 [4]에서 사용한 분석 방법에서는 추정하고자 하는 확률 값이 0 또는 1에 근접하지 않고 표본 크기가 200 이상일 경우 Wald 신뢰구간에 대한 분석은 미비한 실정이다. 또한 [4]에서는 [3]에서 추천한 Agresti-Coull 신뢰구간은 비교 대상에 포함되지 않았으며, 따라서 단측검정을 사용할 경우에 Agresti-Coull 신뢰구간에 대한 분석이 필요하다.

본 논문에서는 확률 값 시험 특성을 고려하여 40에서 100,000까지의 다양한 크기의 표본에 대해 신뢰구간에 대한 비교 분석을 수행한다. 기존에 가장 널리 사용되어 온 Wald 신뢰구간과 Agresti-Coull 신뢰구간을 분석 대상으로 하며, 기존의 커버리지 확률과 신뢰구간의 길이를 기반으로 한 분석과는 달리 해당 신뢰구간을 확률 값 시험에 사용했을 때 측정되는 P_{EP}와 P_{NEF}를 기반으로 하여 비교 분석을 수행한다[1]. 또한 확률 값이 올바른지를 판단하기 위한 양측검정뿐만 아니라 확률 값이 기준 확률 이상인지 여부를 판단하기 위한 단측검정을 사용하는 경우에 대해서도 비교 분석을 수행한다.

본 논문의 구성은 다음과 같다. 2절에서는 구간 추정 기법에 대해 설명한다. 3절에서는 구간 추정 기법을 이용한 확률 값 시험 방법에 대해 설명하며, 4절에서는 양측검정을 사용하는 경우에 대한 신뢰구간 비교 분석을 수행한다. 5절에서는 단측검정을 사용하는 경우에 대한 신뢰 구간 비교 분석을 수행하며, 6절에서는 결론을 맺는다.

II. 구간 추정 기법

일반적으로 널리 알려진 구간 추정 기법은 Central Limit Theorem (CLT), 즉 많은 수의 동일한 분포 함수 (distribution function)을 가지는 독립적인 랜덤 변수의 합은 정규 분포 (normal distribution)에 가까워진다는 이론을 기반으로 한 Wald 신뢰구간이다 [2]. 모집단의 분포 함수가 확률 값 p를 파라미터로 가지는 이항 분포 (binomial distribution)라 가정하자. 모집단으로부터 얻은 크기가 n인 표본에 대해 n의 값이 충분히 크다면, p에 대해 $100(1-\alpha)\%$의 신뢰수준을 가지는 Wald 신뢰구간은 다음과 같이 주어진다.

$$\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \cdots\cdots (1)$$

이 때 $\hat{p}=X/n$이고 X는 표본에서 확률 값 p을 가지는 이벤트가 관찰된 횟수이며, $z_{\alpha/2}$는 표준 정규(unit normal) 랜덤변수 Z에 대해서 $P\{|Z|<z_{\alpha/2}\}=1-\alpha$를 만족한다.

비록 Wald 신뢰구간이 널리 사용되어 왔지만 p 값이 0 또는 1에 근접할 경우 낮은 커버리지 확률을 제공하는 것으로 잘 알려져 있다. p 값이 0 또는 1에 근접하지 않더라도 n 값이 충분히 크지 않을 경우 낮은 커버리지 확률을 제공하며, p 값이 0 또는 1에 근접하지 않고 n 값이 충분히 크다 하더라도 때때로 낮은 커버리지 확률을 제공하는 경우가 존재한다 [3].

[3]에서는 다양한 종류의 신뢰 구간에 대해 비교 분석을 수행하였으며, Agresti-Coull 신뢰구간을 포함한 몇 개의 신뢰 구간을 사용할 것을 권장하였다. Agresti-Coull 신뢰구간은 Wald 신뢰구간과 유사한 형태를 가지기 때문에 사용하기가 편리하다는 장점을 가진다. p에 대해 $100(1-\alpha)\%$의 신뢰수준을 가지는 Agresti-Coull 신뢰구간은 다음과 같이 주어진다.

$$\tilde{p}-\kappa\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}<p<\tilde{p}+\kappa\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} \quad \cdots\cdots\cdots\cdots (2)$$

이 때 $\tilde{p}=\tilde{X}/\tilde{n}$, $\tilde{X}=X+\kappa^2/2$, $\tilde{n}=n+\kappa^2$이며, κ 는 표준 정규 랜덤변수 Z에 대해서 $P\{|Z|<\kappa\}=1-\alpha$를 만족한다. 단, $\alpha=0.05$일 경우 κ는 1.96 대신 2를 사용한다. Agresti-Coull 신뢰구간은 p가 0또는 1에 매우 근접한 경우라도 매우 높은 커버리지 확률을 제공하는 장점이 있으며, 표본의 크기가 작은 경우라도 높은 커버리지 확률을 제공한다.

III. 확률 값 시험 방법

PFSM의 한 상태에서 입력 a에 대해 p와 $1-p$의 확률 값을 가지고 x와 y를 출력으로 내는 천이 t_1과 t_2가 각각 존재한다고 가정하자. 해당 상태에서 입력 a가 주어졌을 때 t_1과 t_2 중 하나를 선택하여 실행하

1) P_{EP}는 커버리지 확률과 동일하다.

는 행위는 하나의 실험(experiment) 또는 시도(trial)라 한다. n번의 독립적인 실험, 즉 입력 a을 n번 반복하여 제공하고 관찰된 데이터는 하나의 표본이 되며, 이 때 n은 표본의 크기가 된다. X가 t_1이 선택되어 실행된 횟수를 나타낸다면 X는 (n,p)를 파라미터로 가지는 이항 랜덤 변수가 되며, 구간 추정 기법을 사용하여 p 값을 일정 수준의 신뢰수준을 가지는 구간으로 예측할 수 있다.

[7]에서는 입력을 n번 반복하여 관찰된 데이터로부터 구현제품의 확률 값이 올바른 지 여부를 시험하기 위해 다음과 같은 기준을 사용하였다.

$$\begin{cases} \text{성공} & \bar{p}-d \le p \le \bar{p}+d \\ \text{실패} & \text{그 밖의 경우} \end{cases} \quad \cdots\cdots (3)$$

이 때 d는 2절에서 구한 신뢰 구간 길이의 1/2이고, Wald 신뢰구간을 사용할 경우는 $\bar{p}=\hat{p}$, Agresti-Coull 신뢰구간을 사용할 경우는 $\bar{p}=\tilde{p}$이다. 시험 대상의 확률 값이 p'이고 기준 확률 값이 p라 하자. 95%의 신뢰수준을 가지는 p'에 대한 신뢰 구간의 의미는 다음과 같다: *해당 신뢰 구간을 얻기 위해 사용된 기법은 해당 기법을 반복하여 신뢰 구간을 얻었을 때 95% 정도는 p'을 포함한다* [2]. 따라서 95%의 신뢰수준을 가지고 (3)을 기반으로 확률 값에 대한 시험을 수행하면 사용된 신뢰구간의 커버리지 확률 값이 명목수준인 95% 이상일 경우 P_{EP}를 95% 정도로 보장할 수 있고, $|p-p'|>2d$인 잘못된 구현 제품에 대해 P_{NEF}를 97.5% 정도로 보장할 수 있다.

또한 [8]에서는 구현 시스템의 통화 성공률 확률 값이 성능 요구 사항을 만족하는지 여부를 판단하기 위해 다음과 같은 기준을 사용하였다.

$$\begin{cases} \text{성공} & p \le \bar{p}+d \\ \text{실패} & \text{그 밖의 경우} \end{cases} \quad \cdots\cdots (4)$$

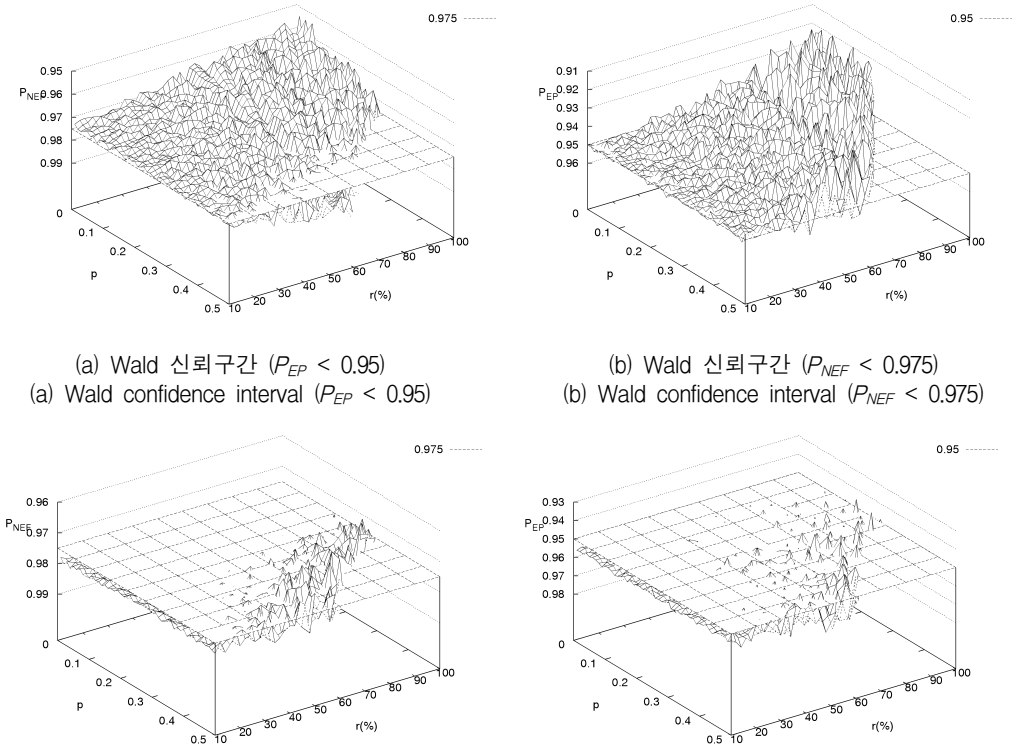
식 (3)과 (4)의 차이는 식 (3)에서는 시험 대상의 확률 값이 기준 확률 값과 유사한지 여부를 판단해야 하므로 양측검정을 사용한 반면 식 (4)에서는 시험 대상의 확률 값이 기준 확률 값 이상인지를 판단해야 하므로 단측검정을 사용한 차이점이 있다. 95%의 신뢰수준을 가지고 (4)를 기반으로 확률 값에 대한 시험을 수행하면 사용된 신뢰구간의 커버리지 확률 값이 명목수준을 보장할 경우 $p' \ge p$ 인 구현 시스템에 대해 P_{EP}를 97.5% 정도로 보장할 수 있고, $p' < p-2d$ 인 잘못된 구현 시스템에 대해 P_{NEF}를 97.5% 정도로 보장할 수 있다.

Ⅳ. 양측 검정 사용 시 신뢰 구간 비교

본 절에서는 식 (3)에서 주어진 것과 같이 확률 값 시험에 양측검정을 사용하는 경우 다양한 확률 값과 다양한 크기의 표본에 대해 P_{EP}와 P_{NEF}를 기반으로 Wald 신뢰구간과 Agresti-Coull 신뢰구간을 비교 분석한다. 비교 분석에서는 편의상 95%의 신뢰수준을 고려하며, 따라서 시험을 통해 95% 정도의 P_{EP}와 97.5% 정도의 P_{NEF}가 보장되어야 한다. 실험을 위해 고려한 PFSM은 하나의 상태를 가지며 입력 a에 대해 p와 $(1-p)$의 확률 값을 가지고 x와 y를 출력으로 내는 두 개의 천이를 가진다. 시험을 위한 기준 확률 값을 p, 시험 대상의 확률 값을 p'이라 하자. 시험 대상 PFSM이 제공되면 PFSM의 각 천이를 확률 값에 따라 실행하는 PFSM 시뮬레이터를 구현하였다. PFSM 시뮬레이터에 입력 a을 주어진 표본 크기만큼 반복하여 제공하고 출력을 관찰한다. PFSM 시뮬레이터를 통해 관찰된 데이터를 기반으로 식 (1)과 (2)를 이용하여 p'에 대한 Wald 신뢰 구간과 Agresti-Coull 신뢰구간을 계산하며, 계산된 신뢰구간을 기반으로 식 (3)을 이용하여 성공 또는 실패 판정이 내려진다.

양측 검정을 사용하여 확률 값을 시험하는 경우에는 시험을 통해 구현 제품의 각 확률 값이 기준 확률과 유사한 값을 가지는지 여부를 판단하게 된다. 따라서 의미 있는 시험 결과를 가지기 위해서는 판정에 사용된 신뢰구간의 길이가 추정하고자 하는 확률 값 대비 적절한 크기를 가져야 한다. 시험 대상의 확률 값 p'에 대한 신뢰구간이 구해지면 자동적으로 확률 값 $(1-p')$에 대한 신뢰구간도 구해지게 된다. 따라서 본 실험에서는 신뢰구간의 길이가 $\min(p,1-p')$의 10%에서부터 100%까지의 값을 가지는 경우를 고려한다.

$\min(p,1-p')$ 값 대비 신뢰구간 길이를 r이라 하자. 본 실험에서는 p가 0.01부터 0.5까지의 값을 가지는 경우를 고려하며, 각 p 값에 대해서 r이 10%에서



(a) Wald 신뢰구간 (P_{EP} < 0.95)
(a) Wald confidence interval (P_{EP} < 0.95)

(b) Wald 신뢰구간 (P_{NEF} < 0.975)
(b) Wald confidence interval (P_{NEF} < 0.975)

(c) Agresti-Coull 신뢰구간 (P_{EP} < 0.95)
(c) Agresti-Coull confidence interval (P_{EP} < 0.95)

(d) Agresti-Coull 신뢰구간 (P_{NEF} < 0.975)
(d) Agresti-Coull confidence interval (P_{NEF} < 0.975)

그림 1. 양측검정 사용 시 신뢰구간 비교
Fig. 1 Comparison of two-sided confidence intervals

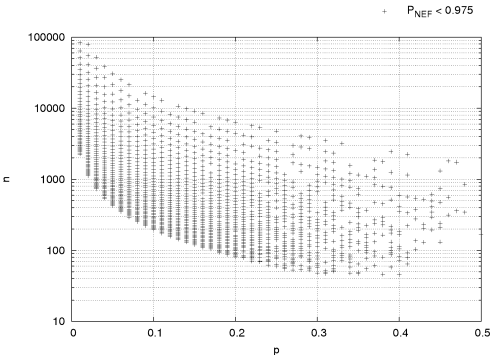
100%까지의 값을 가지도록 표본의 크기를 정하였다. P_{EP}의 경우 $p' = p$이므로 식 (1)과 (2)를 이용하여 주어진 신뢰구간의 길이로부터 표본의 크기 n를 계산할 수 있다. P_{NEF}의 경우 $p' = p + 2d$인 잘못된 구현 제품을 고려하며, [7]에서 제안된 것과 같이 다음에 주어진 조건을 이용하여 표본의 크기 n를 계산한다.

$$n > \left(\frac{\kappa}{d}\right)^2 \check{p}(1-\check{p}) - \kappa^2 \quad \cdots\cdots (5)$$

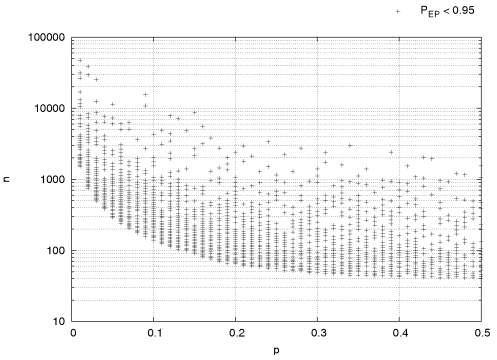
이 때 $\check{p} = p + d$이다. [2]에서는 Wald 신뢰구간을 사용하기에 적합한 표본의 크기를 $np(1-p) \geq 10$으로 추천하였으며, [3]에서는 $n \geq 40$인 경우 Agresti-Coull 신뢰구간을 사용할 것을 권장하였다[2]. 따라서 본 실험에서는 Wald 신뢰구간을 사용했을 경우 표본의 크기가 $np(1-p) < 10$인 경우에는 고려 대상에서 제외하며 Agresti-Coull 신뢰구간을 사용했을 경우는 $n < 40$인 경우는 고려 대상에서 제외한다. 하나의 p와 r의 조합에 대해 100,000번 실험을 반복하였으며, 각 신뢰구간에 대해 P_{EP}와 P_{NEF}를 측정하였다. 그림 1에 실험 결과가 나와 있다.

3절에서 설명한 것처럼 식 (3)을 이용하여 확률 값에 대한 시험을 수행할 경우 95% 정도의 P_{EP}와 97.5% 정도의 P_{NEF}를 보장할 수 있다. 하지만 그림 1에서 보듯이 Wald 신뢰구간의 경우 상당히 많은 p 와 r의 조합에 대해 P_{EP}와 P_{NEF}가 목표 된 확률 값을 만족하지

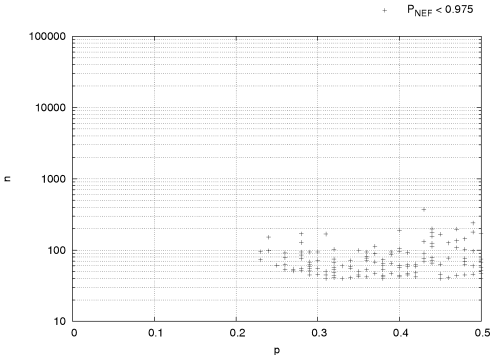
2) $n < 40$ 인 경우 Agresti-Coull 신뢰구간은 높은 커버리지 확률을 제공하지만 상대적으로 긴 신뢰구간을 제공하는 단점이 있다.



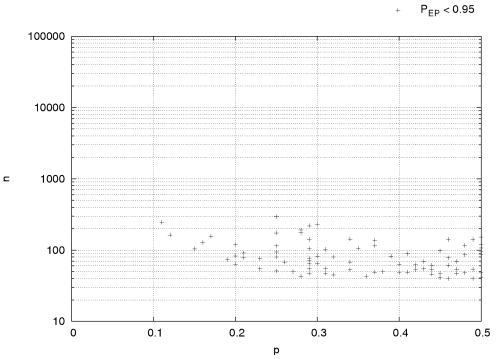
(a) Wald 신뢰구간 (P_{EP} < 0.95)
(a) Wald confidence interval (P_{EP} < 0.95)



(b) Wald 신뢰구간 (P_{NEF} < 0.975)
(b) Wald confidence interval (P_{NEF} < 0.975)



(c) Agresti-Coull 신뢰구간 (P_{EP} < 0.95)
(c) Agresti-Coull confidence interval (P_{EP} < 0.95)



(d) Agresti-Coull 신뢰구간 (P_{NEF} < 0.975)
(d) Agresti-Coull confidence interval (P_{NEF} < 0.975)

그림 2. 양측검정 사용 시 표본 크기에 따른 신뢰구간 비교
Fig. 2 Comparison of two-sided confidence intervals with sample sizes

못하고 있는 것을 볼 수 있다. 신뢰구간의 길이가 길어질수록, 즉 표본의 크기가 작을수록 확률 값의 오류가 커지는 것을 볼 수 있으며, P_{NEF}의 경우 확률값 p가 작아질수록 오류 값이 커지는 것을 볼 수 있다. Agresti-Coull 신뢰구간의 경우 P_{EP}와 P_{NEF} 모두 Wald 신뢰구간에 비해 상대적으로 우수한 결과를 보이고 있으나 Agresti-Coull 신뢰구간 역시 적지 않은 수의 p와 r의 조합에 대해 P_{EP}와 P_{NEF}가 목표 된 확률 값을 만족하지 못하고 있는 것을 볼 수 있다.

[7]에서는 확률 값 시험에 Agresti-Coull 신뢰구간을 사용했을 경우 우수한 P_{EP}와 P_{NEF}를 보장하는 p와 r의 조합들을 제시하고 이를 사용할 것을 권장하였다. 이를 위해 실험을 통해 측정된 P_{EP}와 P_{NEF} 값이 각각 95%와 97.5% 이상일 경우 해당 p와 r의 조합은 우수하다고 판단하였다. 하지만 이러한 판단 기준은 [8]에서 언급된 것과 같이 측정된 P_{EP}와 P_{NEF} 값이 통계적 특성을 가지는 것을 고려하지 않음으로 인해 열악한 우수 판정 확률을 보이는 문제가 있다. 하나의 p와 r의 조합에 대해 P_{EP} 또는 P_{NEF}의 측정을 위해 실험을 100,000번 반복하여 얻은 데이터는 크기가 100,000인 하나의 표본이라 할 수 있다. 표본을 기반으로 P_{EP}가 95% 이상의 값을 가지는지 또는 P_{NEF}가 97.5% 이상의 값을 가지는지 여부를 판단해야 하며, 따라서 식 (4)에서 제시된 판단 기준이 사용될 수 있다. 본 실험에서는 시험 대상 확률 값이 1에 가까울 경우 우수한 성능을 보이는 Agresti-Coull 신뢰구간을 이용하여 각 p와 r 조합에 대해 측정된 확률 값을 기반으로 P_{EP}와 P_{NEF}가 각각 95%와 97.5% 이상의 확률 값을 가지는지 여부를 판단하였다. 그림 2에는 각 확률 값 p에 대해 P_{EP}와 P_{NEF}가 각각 95%

와 97.5% 미만의 확률 값을 가지는 경우가 나와 있다. 그래프의 y축에는 표본의 크기와 실험 결과와의 상관관계를 살펴보기 위해 r 대신 표본의 크기 n을 나타내었다.

그림에서 보듯이 Wald 신뢰구간의 경우 표본의 크기가 충분히 큼에도 불구하고 P_{EP}와 P_{NEF}가 각각 95%와 97.5% 미만인 경우가 상당히 많음을 알 수 있다. 약 36% 정도의 p와 r의 조합에서 P_{EP}가 95% 이상 확률 값을 가지는 것으로 판단되었으며, 약 31% 정도의 조합에서 P_{NEF}가 97.5% 이상 확률 값을 가지는 것으로 판단되었다. P_{EP}의 경우 모든 확률 값에서 비슷한 수준으로 좋지 않은 결과를 보여주고 있으며, 대부분의 확률 값에서 표본의 크기가 1,000 이상인 경우에도 적지 않게 P_{EP}가 95%보다 작은 확률 값을 가지는 경우가 존재함을 알 수 있다. P_{NEF}의 경우 확률 값이 0.5에 근접할 경우에는 양호한 결과를 보여주고 있으나 0.4 이하일 경우 확률 값이 작아질수록 열악한 결과를 보여주고 있음을 알 수 있다. 특히 확률 값이 0에 가까운 경우에는 100,000에 가까운 표본 크기를 가지더라도 P_{NEF}가 97.5%보다 작은 확률 값을 가지는 것으로 나타났다.

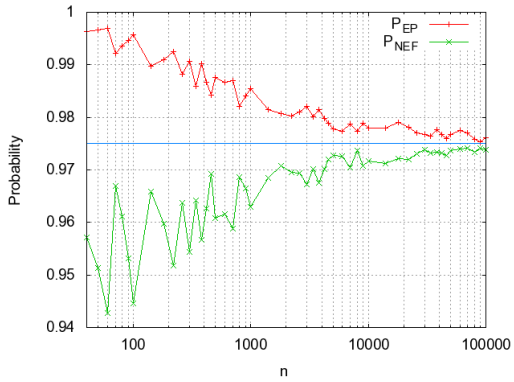
Agresti-Coull 신뢰구간의 경우 약 96% 정도의 p와 r의 조합에서 P_{EP}가 95% 이상의 값을 가지는 것으로 판단되었으며, 약 94% 정도의 조합에서 P_{NEF}가 97.5% 이상의 확률 값을 가지는 것으로 판단되었다. 확률 값이 0.2 이하인 경우에 매우 우수한 결과를 보여주고 있으며, 표본의 크기가 약 200 이상인 경우 모든 확률 값에서 P_{EP}와 P_{NEF} 모두 우수한 결과를 보여주고 있음을 알 수 있다.

Ⅴ. 단측검정 사용 시 신뢰 구간 비교

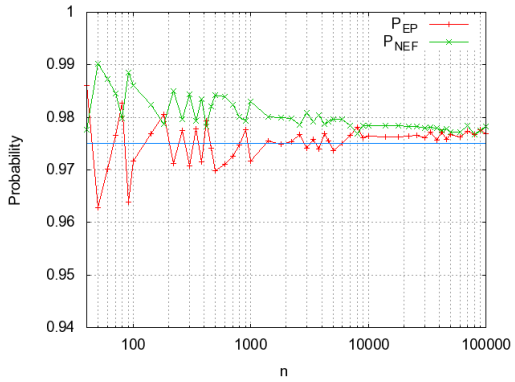
본 절에서는 식 (4)에서 주어진 것과 같이 확률 값 시험에 단측검정을 사용하는 경우 다양한 확률 값과 다양한 크기의 표본에 대해 P_{EP}와 P_{NEF}를 기반으로 Wald 신뢰구간과 Agresti-Coull 신뢰구간을 비교한다. 실험 방법은 기본적으로 성공 또는 실패 판정을 내리기 위한 판정 기준만 다를 뿐 4절에서 설명한 방법과 동일하다. [8]에서와 같이 확률 값 시험에 단측검정을 이용하는 경우는 시험 대상의 확률 값이 요구된 기준 이상인지 여부를 판단하는 경우로 기준 확률 값은 일반적으로 1에 가까운 값을 가진다. 또한 [9]에서와 같이 시험 대상이 되는 시스템이 상용 서비스 중일 수도 있으므로 시스템에 큰 부하를 주기 어려울 수도 있으며, 적은 크기의 표본을 기반으로 판정이 내려져야 하는 경우도 존재한다. 따라서 본 실험에서는 0.8, 0.9, 0.95 세 가지 확률 값을 고려하며, 각 확률 값에 대해 40부터 100,000까지 다양한 표본 크기를 이용하여 비교 분석을 수행한다. P_{NEF}의 경우 4절에서와 같이 $p' = p + 2d$인 잘못된 구현 제품을 고려하였으며, 식 (5)를 이용하여 표본의 크기로부터 d를 계산하였다. 각 확률 값 p와 표본의 크기 n의 조합에 대해 100,000번 실험을 반복하였으며, 각 신뢰구간에 대해 P_{EP}와 P_{NEF}를 측정하였다. 그림 3에 실험 결과가 나와 있다.

3절에서 설명한 것처럼 식 (4)를 이용하여 확률 값에 대한 시험을 수행할 경우 97.5% 정도의 P_{EP}와 97.5% 정도의 P_{NEF}를 보장할 수 있다. P_{EP}의 경우 그림 3에서 보는 것과 같이 작은 크기의 표본에서는 Wald 신뢰구간이 Agresti-Coull 신뢰구간보다 우수한 결과를 보인다. 하지만 Wald 신뢰구간의 경우 확률 값이 0.8인 경우에서 보듯이 표본의 크기가 매우 큼에도 불구하고 (10,000 이상) P_{EP}가 97.5% 보다 작은 경우가 존재한다. Agresti-Coull 신뢰구간의 경우 작은 크기의 표본에 대해서는 P_{EP}가 97.5% 보다 작은 경우가 많이 존재하지만 표본의 크기가 커짐에 따라 0.975와 0.98 사이의 값으로 안정적으로 수렴해 가는 것을 볼 수 있다. P_{NEF}의 경우에는 Agresti- Coull 신뢰구간이 모든 확률 값에서 우수한 결과를 보인다. Wald 신뢰구간을 이용할 경우 P_{NEF}가 97.5% 이상인 경우가 드물게 존재하지만 Agresti- Coull 신뢰구간을 이용할 경우 대부분의 경우에서 P_{NEF}가 97.5% 보다 높은 값을 보이고 있다.

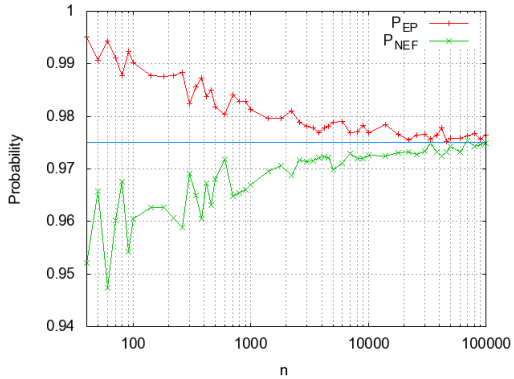
본 논문에서는 단측검정을 사용하여 0.8 이상의 확률 값을 시험할 경우 큰 크기의 표본 (약 1,500 이상)에 대해서는 Agresti-Coull 신뢰구간을 사용할 것을 추천한다. 표본의 크기가 작을 경우에는 시험의 목적에 따라 달라지며, 낮은 P_{NEF} 값을 가지더라도 우수한 P_{EP}가 보장되어야 하는 경우에는 Wald 신뢰구간을, 상대적으로 낮은 P_{EP} 값을 가지더라도 우수한 P_{NEF}가 보장되어야 하는 경우에는 Agresti-Coull 신



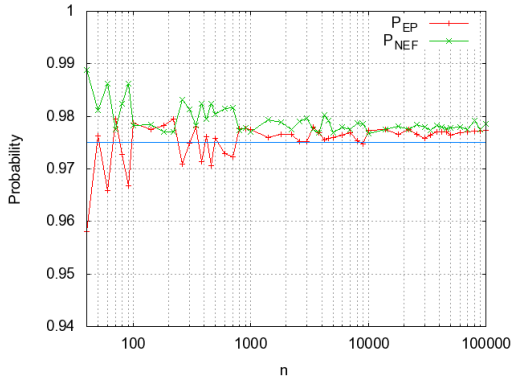
(a) Wald 신뢰구간 $(p=0.95)$
(a) Wald confidence interval $(p=0.95)$



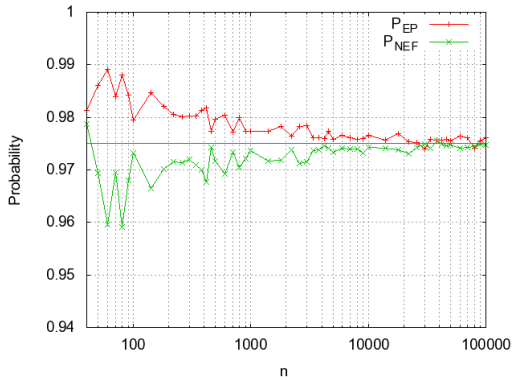
(b) Agresti-Coull 신뢰구간 $(p=0.95)$
(b) Agresti-Coull confidence interval $(p=0.95)$



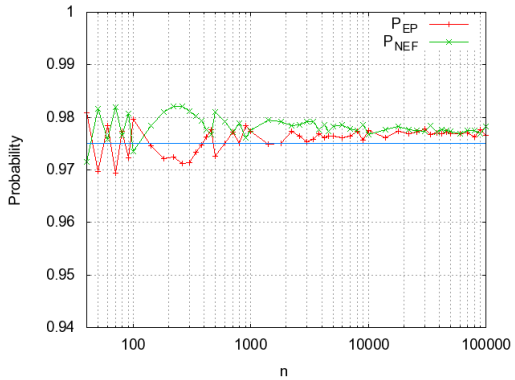
(c) Wald 신뢰구간 $(p=0.9)$
(c) Wald confidence interval $(p=0.9)$



(d) Agresti-Coull 신뢰구간 $(p=0.9)$
(d) Agresti-Coull confidence interval $(p=0.9)$



(e) Wald 신뢰구간 $(p=0.8)$
(e) Wald confidence interval $(p=0.8)$



(f) Agresti-Coull 신뢰구간 $(p=0.8)$
(f) Agresti-Coull confidence interval $(p=0.8)$

그림 3. 단측검정 사용 시 신뢰구간 비교
Fig. 3 Comparison of one-sided confidence intervals

뢰구간을 사용할 것을 추천한다. 기준 확률 값이 0.8 미만인 경우에는 [4]에서 보인 것과 같이 본 실험에서 제시된 결과와는 다른 결과를 보일 수 있다. 따라서 향후 다양한 확률 값에 대한 보다 일반적인 비교 분석이 수행되어야 한다.

Ⅵ. 결 론

본 논문에서는 신뢰구간이 확률 값 시험에 사용되는 경우의 특성을 고려하여 40에서 100,000까지의 다양한 크기의 표본에 대해 Wald 신뢰구간과 Agresti-Coull 신뢰구간을 비교 분석하였다. 올바른 구현 제품이 시험을 통과할 확률(P_{EP})과 잘못된 구현 제품이 시험을 통과하지 못할 확률(P_{NEF})을 기반으로 하여 비교 분석을 수행하였으며, 확률 값이 올바른지를 판단하기 위한 양측검정뿐만 아니라 확률 값이 기준 확률 값 이상인지 여부를 판단하기 위한 단측검정을 사용할 경우에 대해서도 비교 분석을 수행하였다. 비교 분석 결과 확률 값 시험에 양측검정이 사용되었을 경우 Wald 신뢰구간은 작은 크기의 표본뿐만 아니라 큰 크기의 표본에 대해서도 열악한 결과를 보임을 알 수 있었으며, Agresti-Coull 신뢰구간의 경우 표본의 크기가 약 200 이상인 경우에는 매우 우수한 결과를 보여주었다. 0.8 이상인 확률 값에 대해 단측검정을 사용하여 확률 값을 시험할 경우에는 큰 크기의 표본 (약 1,500 이상)에 대해서는 Agresti-Coull 신뢰구간이 우수한 결과를 보였으며, 작은 크기의 표본에 대해서는 P_{EP}의 경우 Wald 신뢰구간이, P_{NEF}의 경우 Agresti-Coull 신뢰구간이 보다 우수한 결과를 보임을 알 수 있었다.

단측검정을 사용하여 확률값을 시험할 경우 표본의 크기가 작을 경우에는 Agresti-Coull 신뢰구간에서도 P_{EP}가 97.5%보다 작은 경우가 다수 존재한다. 따라서 향후 연구방향으로는 단측검정을 사용하여 확률 값을 시험하는 경우 Jeffreys 신뢰구간을 포함한 보다 다양한 신뢰구간에 대한 비교 분석이 수행되어야 한다. 또한 0.8 미만의 확률 값에 대해 단측검정을 사용할 경우 다양한 신뢰구간에 대한 비교 분석도 수행되어야 한다.

참고 문헌

[1] O. Markowitch, Y. Roggeman. Probabilistic non-repudiation without trusted third party. Second Conference on Security in Communication Networks 1999, Amalfi, Italy.

[2] S.M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, Johns Wilely & Sons, New York, 1987.

[3] L. D. Brown, T. T. Cai, A. DasGupta. "Interval Estimation for a Binomial Proportion," Statistical Science, Vol. 16, No. 2, pp. 101-117, 2001.

[4] T. T. Cai, "One-sided confidence intervals in discrete distributions", Journal of Statistical Planning and Inference, Vol. 131, No. 1, pp. 63-88, 2005.

[5] A. Agresti, B. Coull, "Approximate is better than 'exact' for interval estimation of binomial proportions", The American Statistician, Vol. 52, No. 2, pp. 119-126, 1998.

[6] 류제복, "이항신뢰구간에 대한 소고", 한국통계학회논문집, 제16권 제5호, pp. 731-743, 2009.

[7] Iksoon Hwang, Ana Cavalli, Testing a probabilistic FSM using interval estimation, Computer Networks, Vol. 54, No. 7 pp. 1108-1125, 2010.

[8] 황익순, 박재성, "구간 추정 기법을 이용한 이동통신 시스템의 통화 성공률 시험 방법", Technical report, 수원대학교, 2010.

[9] 방송통신위원회 이용자보호국, 2009년도 통신서비스 품질 평가 결과, 2010. URL http://www.wiseuser.go.kr/quality/main.do

저자 소개

황익순(Ik-soon Hwang)

1995년 2월 연세대학교 전자공학과 졸업 (공학사)

1997년 2월 연세대학교 대학원 전자공학과 졸업(공학석사)

2001년 2월 연세대학교 대학원 전기전자공학과 졸업 (공학박사)

프랑스 TELECOM SudParis 연구원

※ 주 관심분야 : Model-based testing, 통신 프로토콜 및 서비스