The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.485
32(4), 485–502

Theoretical statistics education using mathematical softwares

Geung-Hee Lee[a1]

[a]Department of Data Science and Statistics, Korea National Open University

(Received May 17, 2019; Revised June 6, 2019; Accepted June 13, 2019)

Abstract

Theoretical statistics is a calculus based course. However, there are limitations to learn theoretical statistics when students do not know enough calculus techniques. Mathematical softwares (computer algebra systems) that enable calculus manipulations help students understand statistical concepts, by avoiding the difficulties of calculus. In this paper, we introduce mathematical software such as Maxima and Wolfram Alpha. To foster statistical concepts in theoretical statistics education, we present three examples that consist of mathematical derivations using wxMaxima and statistical simulations using R.

Keywords: mathematical statistics, statistical inference, CAS, Maxima, Wolfram Alpha, R

1. 머리말

우리는 데이터 홍수 시대에 살고 있다. 또한 데이터를 저장, 분석할 수 있는 컴퓨팅 환경과 관련 알고리즘이 급속히 발전하면서 데이터 기반 분석 및 의사결정이 확산되고 있다. 이로 인해 데이터 분석의 원리인 통계적 추론과 관련된 이론통계학(수리통계학)에 대한 관심도 같이 커지고 있다. 이론통계학은 통계학의 원리를 수학을 이용하여 도출하거나 확장하는 것을 배우는 교과목이고, 국내외 통계학과의 학부, 대학원 교육의 중심 교과목이다. 이론통계학 교육은 교수가 수학 증명 및 연산을 강의하고 학생들이 연습문제를 반복적으로 풀어서 통계학의 원리를 배우는 과정으로 진행되고 있다. 학생들이 미적분 등 수학 지식이 부족한 경우 이론통계학의 문제풀이를 통해 통계학의 원리를 습득하기 어렵다. 그런데 이론통계학에서 수학은 통계학 원리를 표현하는 형식 (과학의 언어)이여서 중심극한정리 증명, 최대가능도 추정량 도출 등 통계학 원리를 수학으로 증명하거나 도출했다고 해서 통계학 원리 자체를 이해했다고 보기는 어렵다. 따라서 이론통계학 교육에서 학생들이 시뮬레이션 등을 통해 통계학 원리를 이해하고 수학을 이용한 증명, 문제풀이를 통해 통계학 원리를 확인·확장할 수 있는 방안을 모색할 필요가 있다. 이러한 이론통계학 교육은 학생들이 데이터 분석을 기계적으로 하지 않고 통계학 원리에 기반하여 유연하게 할 수 있게 한다.

이론통계학 교육에서 수학의 장벽을 극복할 수 있는 한 방법은 소프트웨어를 이용하는 것이다. 소프트웨어는 이론적 결과를 시각적으로 체감하고 강화하여 학생들의 학습 성과를 높인다 (Hunter, 2005;

The work was supported by the Korea National Open University in 2017.
[1]Department of Data Science and Statistics, Korea National Open University, 86, Daehak-ro, Jongno-Gu, Seoul 03087, Korea. E-mail: geunghee@knou.ac.kr

Table 2.1. List of computer algebra systems

CAS Name	Creator	Release Year	Notes
Maxima	W. F. Schelter	1998	Free, Android
SageMath	W. A. Stein	2005	Free
Mathematica	Wolfram Research	1988	Commercial CAS
Maple	University of Waterloo	1984	Commercial CAS
Matlab	Math Works	2008	Commercial CAS
Wolfram Alpha	Wolfram Research	2009	Online CAS

주) source: wikipedia (List of computer algebra systems)

Green과 Blankenship, 2015). 이러한 연구들을 바탕으로 ASA는 통계교육 평가·강의 지침(Guidelines for Assessment and Instruction in Statistics Education)을 작성하였는데, 여기서 ASA는 소프트웨어를 도입하여 통계학을 개념적으로 교육할 것을 권고하고 있다 (Carver 등, 2016).

이론통계학 교육에서 이용할 수 있는 소프트웨어는 시뮬레이션 소프트웨어와 수학 연산 소프트웨어로 구분된다. 시뮬레이션 소프트웨어는 통계적 추론의 추상적 개념을 시각화하여 이해할 수 있게 하는데 JAVA/Javascript 기반 웹기반 GUI 앱과 R Shiny 기반 웹기반 GUI 앱이 있다 (JAVA/Javascript 기반 웹기반 GUI 앱은 http://www.rossmanchance.com/applets/를, R Shiny 기반 웹기반 앱은 https://statistics.calpoly.edu/shiny를 참조하면 된다) (Doi 등, 2016). 시뮬레이션은 중요한 통계적 명제를 확인하거나 증명 또는 연산 도출의 대안이 될 수 있어서 표본분포, 유의확률, 신뢰구간 등 통계적 추론의 개념을 시각적으로 탐색하는데 도움을 준다 (Moore, 1997; Rossman과 Chance, 1999; Horton 등, 2004; Jang 2009). 수리통계학 교재도 R 프로그램을 이용하여 통계적 추론의 개념을 보강하는 방향으로 작성되고 있다 (Kim, 2011; Kang과 Park, 2015).

수학 연산 소프트웨어는 수학 수식 연산을 빠르고 정확하게 할 수 있게 하여 통계학 개념을 수리적으로 추상화하고 확장할 수 있게 한다. 수학 기호 연산이 가능한 수학 소프트웨어 Computer Algebra System (CAS)로는 Maxima, SageMath, Mathematica, Maple, Matlab (Symbolic Math Toolbox), Wolfram Alpha 등이 있다. 선진국과 우리나라는 수학교육에서 SageMath, Geogebra, Maple, Mathematica 등을 이용하여 수학연산을 시각화하고 도구를 이용한 연산을 시도하고 있다 (Lee와 Park, 2015). 본고에서는 CAS 중 무료로 이용할 수 있는 Maxima, Wolfram Alpha를 소개하고, 이를 이용한 문제풀이와 R을 이용한 시뮬레이션을 결합하여 이론통계학 교육을 개선할 수 있는 방안을 모색하고자 한다. 본고의 구성은 제2장에서 수학 프로그램인 Maxima, Wolfram Alpha 등을 소개하고, 제3장에서 R을 이용한 시뮬레이션과 wxMaxima를 활용한 수학 기반 문제풀이를 결합한 이론통계학 교육 사례를 정리한다. 제4장에서는 이론통계학 교육 방향에 대해 정리한다.

2. 수학소프트웨어: Maxima와 Wolfram Alpha

수학 소프트웨어는 미분, 적분, 행렬연산 등 수학 연산을 할 수 있는 프로그램이며 CAS라고 부른다. CAS는 1970년대 초에 개발되기 시작되었고 1세대 소프트웨어로는 Reduce, Derive와 Macsyma 등이 있다. 주요 CAS는 Table 2.1에 정리되어 있는데 이를 보면 공개프로그램인 Maxima, SageMath와 유료 프로그램인 Mathematica, Maple, Wolfram Alpha, Matlab (Symbolic Math Toolbox) 등으로 구분된다. Wolfram Alpha의 경우 인터넷 사이트를 통해 무료로 이용할 수 있으나 전문적 이용을 위해서는 유료 구독 서비스에 가입해야 한다. R 프로그램의 경우 mosaicCalc 패키지나 Yacas를 연결한 Ryacas 패키지를 이용하여 수학연산을 할 수 있으나, 앞서의 CAS 프로그램에 비해 그 성능이 제한적이다. 수

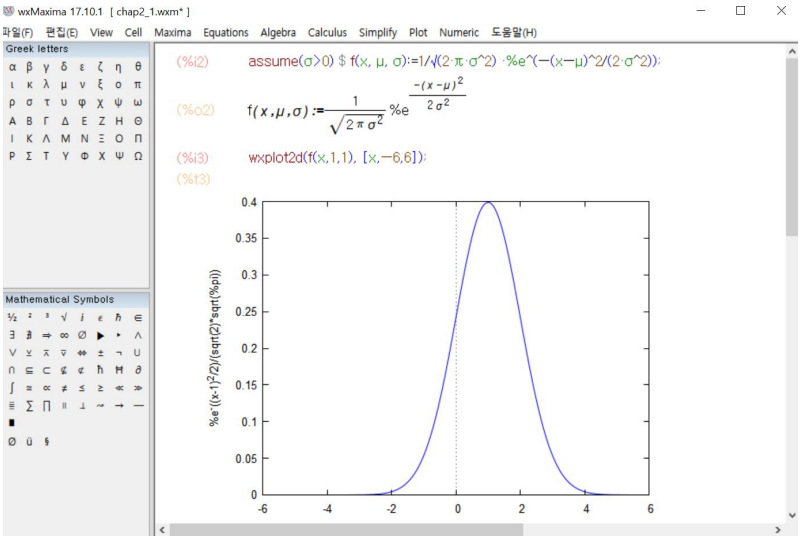


Figure 2.1. Normal distribution using wxMaxima.

학 소프트웨어를 이용한 이론 통계학 교육에 적용한 연구는 주로 Maple과 Mathematica를 이용하여 진행되었는데, 이에 대한 연구로는 Baglivo (1995), Berger (1998), Rose와 Smith (2000) 등이 있다. CAS는 수학 연산을 쉽게 접근할 수 있게 하고 연산 시간을 줄여서 유사한 문제를 반복적으로 학습할 수 있게 한다. 그러나 학생들이 수학과 더불어 CAS 언어를 추가적으로 배워야 하며, CAS에 지나치게 의존하여 수학적 개념을 놓칠 수 있다. 또한 복잡한 수학 문제의 경우 CAS만으로 문제를 일괄적으로 풀 수 없다 (Kumar과 Kumaresan, 2008). 따라서 CAS는 계산기와 같이 수학적 풀이의 보조적, 시각적 도구로 이용할 필요가 있다. 이 절에서는 무료로 이용할 수 있는 Maxima와 Wolfram Alpha를 소개하고, R에서 수학연산을 할 수 있는 방법을 살펴본다.

2.1. Maxima의 소개

Maxima는 W. F. Schelter가 MIT에서 개발했던 Macsyma 프로그램 1982년 버전을 바탕으로 1998년 다시 만든 수학 소프트웨어이다. Maxima의 활용도를 높이기 위해서 다양한 패키지가 있는데 통계 관련 패키지로는 stat, distrib 등이 있다. Maxima를 보다 쉽게 이용하기 위해서는 wxWidgets 기반 GUI 형태의 wxMaxima를 이용할 필요가 있다 (wxMaxima는 http://andrejv.github.io/wxmaxima/에서 내려 받을 수 있다). wxMaxima의 화면은 Figure 2.1과 같다. 이 화면에서 입력선 `(%i)`에 명령어를 입력한 후 한줄 씩 `shift+Enter` 또는 `Crtl+Enter`를 클릭하거나 상단 메뉴의 `Cell`을 선택하여 프로그램을 수행하며, 수행된 결과는 `(%o)`에 나타난다. wxMaxima에는 그리스 문자와 수학 연산자를 수식에 입력할 수 있으며, 텍스트를 목차로 정리할 수 있다는 장점이 있다. wxMaxima 관련 수학 연산 기능과 명령어는 프로그램의 도움말과 Leydold와 Petry (2011)를 참조하면 된다.

Figure 2.1은 정규분포 확률밀도함수를 함수로 정의하고 그래프로 표현한 결과이다. 첫줄 `assume(`$\sigma > 0$`) $`는 σ를 양수로 가정하고 결과로 표시하지 않는 것을 의미한다. `wxplot2d`는 정규분포 확률밀도함수를 그래프를 그리게 하는 함수이다. Figure 2.2는 정규분포 확률밀도함수를 `diff`함수를 이용하여 미분하고, `integrate`함수를 이용하여 정규분포의 기댓값을 구한 것이다. 아울러 적률생성함수 `mgf`를

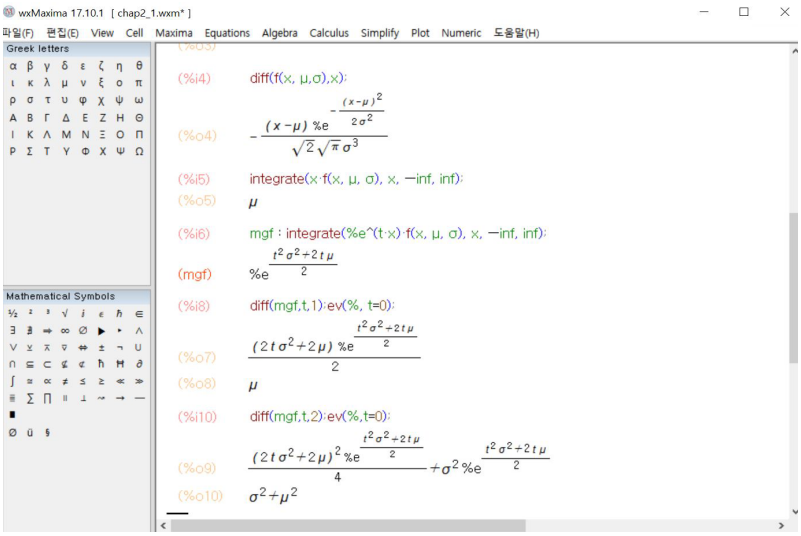
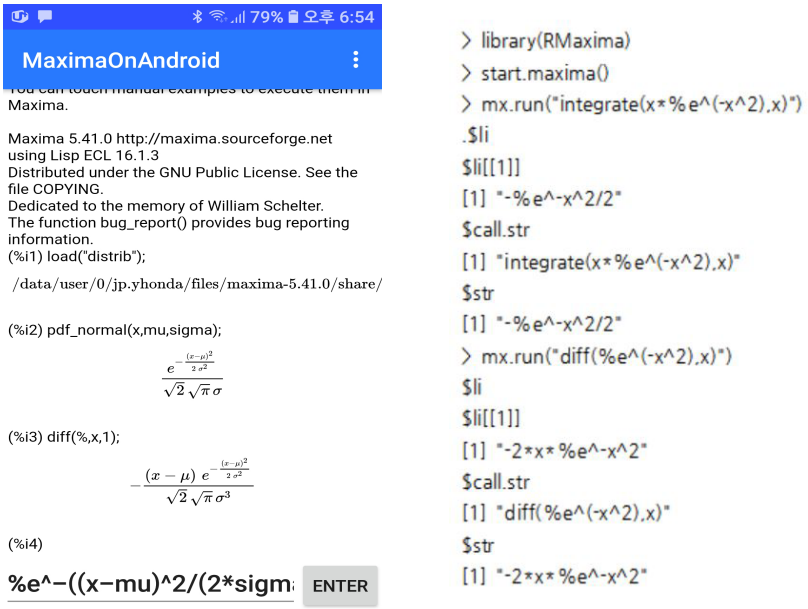


Figure 2.2. Moment generating function using wxMaxima.

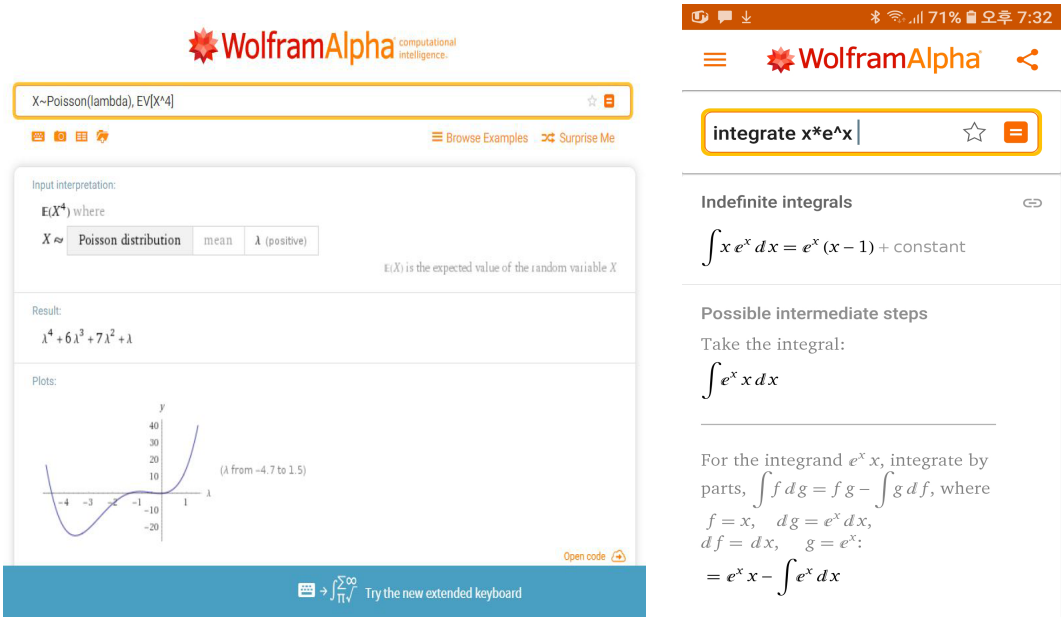


(a) Mobile app (b) RMaxima

Figure 2.3. Maxima mobile app and RMaxima.

`integrate`함수를 이용하여 정의하고, 이를 `diff`함수로 1차, 2차 미분한 후 $t = 0$를 대입한 `ev`함수를 이용하여 1차 적률과 2차 적률을 구한 것이다. 여기서 %는 바로 전에 계산한 수식을 의미한다.

Figure 2.3(a)는 안드로이드 OS에서 사용할 수 있는 Maxima 무료 앱인데 하단에서 명령어를 입력하고 Enter를 클릭하면 수학 연산 결과를 볼 수 있다. 또한 R 환경하에서 RMaxima 패키지



(a) PC (b) Mobile app

Figure 2.4. Wolfram Alpha

```
normal distribution
mgf of normal distribution
d/dt mgf of normal distribution, t=0
d^2/dt^2 mgf of normal distribution, t=0
```

Figure 2.5. Key words of Wolfram Alpha.

를 이용하여 Maxima를 이용할 수 있는데 Figure 2.3(b)와 같다 (RMaxima에 대한 자세한 내용은 https://github.com/skranz/RMaxima를 참조). 그런데 RMaxima로 Maxima의 기능을 충분히 활용하는 데에는 제약이 있다.

2.2. Wolfram Alpha의 소개

Wolfram Alpha는 Mathematica 개발자 Stephen Wolfram이 2009년 개발한 지식 검색엔진으로 수학 연산과 통계연산을 포함한 세상의 지식을 웹(https://www.wolframalpha.com)에서 검색할 수 있다. Wolfram Alpha의 검색창에 의미 있는 수학, 통계학 명령어 등 입력하면 스스로 인식하여 다양한 지식 결과물을 제공한다. 다른 CAS 프로그램과 달리 프로그램을 설치하지 않고 인터넷으로 결과를 찾을 수 있다. Figure 2.4는 각각 Wolfram Alpha를 PC와 모바일 앱(유료)의 검색 결과이다. Figure 2.4(a)를 보면 포아송분포를 따르는 확률변수의 4차 적률을 구한 결과이다. 이를 보면 적률 연산결과와 결과 그래프 등 연산과 관련된 다양한 결과를 볼 수 있다. Figure 2.4(b)는 모바일에서 xe^x를 적분한 결과인데 풀이과정도 살펴볼 수 있다. Figure 2.5는 정규분포, 정규분포의 적률생성함수, 정규분포 적률생성함수를 미분해서 구하는 1차 적률과 2차 적률을 구하는 검색어이다. 이와 같이 Wolfram Alpha를 이용한 통계 연산 검색은 유연하지만 일괄 처리가 어려운 제약이 있다.

```
> ###  mosaicCalc 패키지를 이용한 미분과 적분 ###
> library(mosaicCalc)
> symbolicD(a*x^2 ~ x)
function (x, a = 3)
a * (2 * x)
> symbolicAntiD(a*x^2 ~ x)
a * 1/(3) * x^3 ~ x
```

Figure 2.6. Calculus using R mosaicCalc.

```
> ###  Ryacas 패키지를 이용한 미분과 적분 ###
> library(Ryacas)
> xs <- Sym("xs")
> Integrate(xs^a, xs)
expression(xs^(a + 1)/(a + 1))
> deriv(xs^a, xs)
expression(a * xs^(a - 1))
```

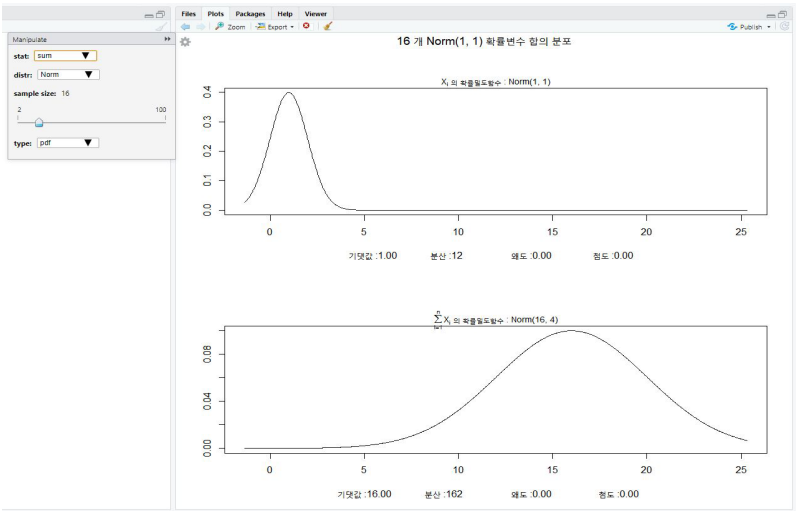
Figure 2.7. Calculus using Ryacas.

```
> ###  distrEx 패키지를 이용한 확률분포의 연산 ###
> library(distrEx)
> X_1 = Norm(0,1); X_2 = Norm(0,1)
> X = X_1 + X_2
> X
Distribution Object of Class: Norm
 mean: 0
 sd: 1.4142135623731
> E(X)
[1] 0
> var(X)
[1] 2
> E(2*X+3)
[1] 3
> var(2*X+3)
[1] 8
```

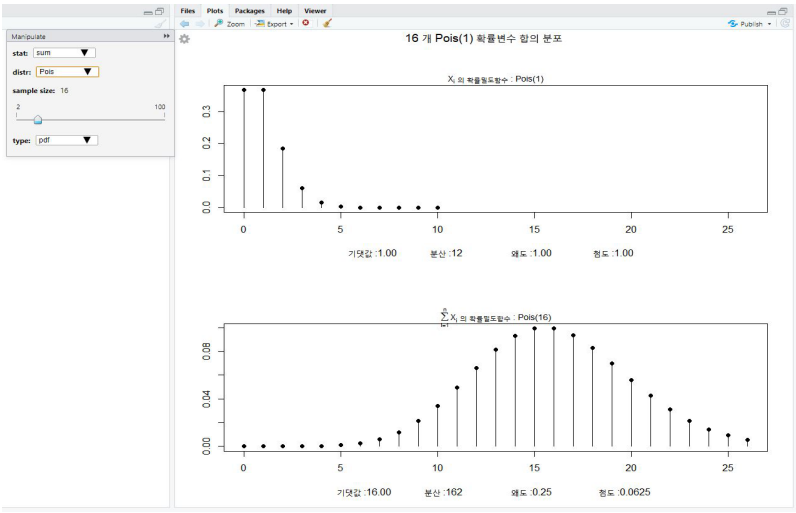
Figure 2.8. Sum of two normal random variables using R distrEx.

2.3. R을 이용한 수학 연산

R에서의 미적분은 기본적으로 D함수와 integrate함수를 통해서 실시할 수 있다. 그러나 integrate 함수는 기호 연산을 할 수 없는 제약이 있다. 미분과 적분의 기호연산을 본격적으로 위해서는 Figure 2.6과 같이 R의 mosaicCalc 패키지의 symbolicD, symbolicAntiD를 이용할 수 있으나 기능적 제약이 있다. 또한 Figure 2.7과 같이 Yacas 수학 프로그램의 인터페이스인 Ryacas 패키지를 이용할 수 있다 (Yacas는 https://yacas.readthedocs.io을 참조). Ryacas 패키지의 기호 연산 기능은 mosaicCalc 패키지보다 다양하나 다른 일반 수학 소프트웨어보다는 제약이 있다. 한편 R의 distEx 패키지로 확률분포의 연산, 기댓값 계산 등을 실시할 수 있다. Figure 2.8은 distEx 패키지를 이용하여 정규분포를 따르는 확률변수들의 합(또는 연산)의 분포를 구하고 이의 기댓값과 분산을 직접 계산한 것이다.



(a) Normal distribution

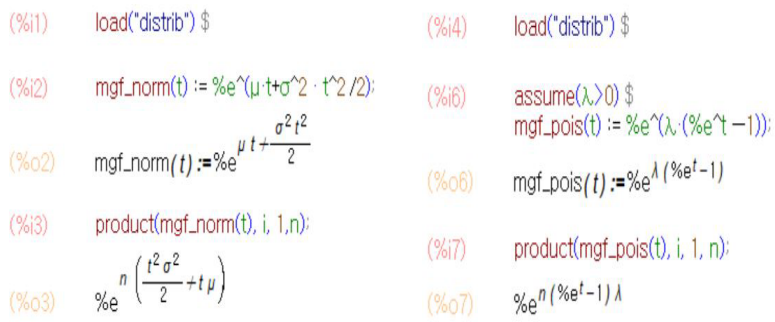


(b) Poisson distribution

Figure 3.1. Distributions of sum of random variables using R.

3. 이론통계학 교육과 수학 소프트웨어의 활용

본 절에서는 이론통계학 교육에서 소프트웨어를 이용한 시뮬레이션과 수학소프트웨어를 이용한 문제풀이를 통해 수학의 배경지식이 적은 학생들이 통계학의 원리를 배우는 방법을 모색하고자 한다. 이를 위해 확률변수 합의 분포의 이해, 이항분포의 포아송 근사, 지수분포 및 포아송분포 모수의 최대가능도추정법 추정을 R을 이용한 시뮬레이션과 wxMaxima를 이용한 수학 증명으로 학습하는 것을 예를 들어 살펴본다. 이 절의 R을 이용한 시뮬레이션은 manipulate 패키지로 작성한 간단한 GUI 프로그램을 이



(a) Normal distribution (b) Poisson distribution

Figure 3.2. Distributions of sum of random variables using wxMaxima.

용하였다. 그 코드는 부록에 있다.

3.1. 확률분포 합의 분포 도출

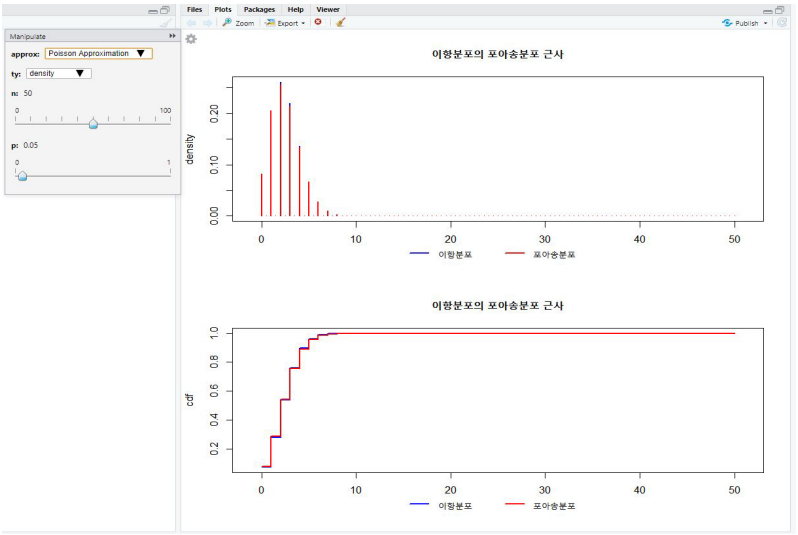
확률변수의 합 또는 평균의 분포는 변수 변환 또는 적률생성함수를 통해 구하는 것이 일반적이다. 이 절에서는 R의 distrEx 패키지를 통해 확률변수의 수가 커지면서 이항분포, 포아송분포, 지수분포, 카이제곱분포, 정규분포를 따르는 확률변수의 합 또는 평균의 분포의 모습을 확률밀도함수 또는 누적분포함수를 통해서 살펴볼 수 있는 GUI 프로그램을 작성하였는데 Figure 3.1과 같다. Figure 3.1을 보면 왼쪽에 통계량(합, 표본평균), 분포, 표본수, 형태(cdf, pdf)를 조정할 수 있는 컨트롤 패널이 있고 오른쪽에 두 개의 그래프가 나타난다. 상단 그래프는 확률변수의 분포 그래프이고, 하단 그래프는 통계량의 그래프이다. 그래프 상단의 제목 부분에 합의 분포가 나오고, 하단에 기댓값, 분산, 왜도와 첨도 값이 나온다. 왼쪽 패널 값을 조정해서 확률분포와 표본수, 확률분포형태에 따라 합(표본)의 분포, 분포 그래프, 기댓값, 분산, 왜도와 첨도의 변화를 볼 수 있다. 확률변수 합의 확률분포를 R의 그래프로 반복해서 살펴보면 합의 확률분포가 표본수, 분포에 따라 어떻게 변하는지를 시각적으로 확인할 수 있다.

앞서의 확률표본 합의 분포의 시각적 결과를 바탕으로 확률변수 합의 분포를 적률생성함수를 이용하여 수학적으로 도출할 수 있고, 이에 대한 wxMaxima 프로그램은 Figure 3.2(a)와 같다. 이를 보면 `(%i2)`에서 `load("distrib")` $는 분포관련 `distrib`패키지를 불러오고, 적률생성함수를 각각 `mgf_norm`과 `mgf_pois`로 지정한다. 확률변수 합의 적률생성함수는 개별 확률변수의 적률생성함수의 곱으로 표현되므로 이를 `product`함수를 이용하여 연산한다. 적률생성함수를 연산한 결과로부터 $N(n\mu, \sigma^2)$, Poisson(λ)의 확률변수 합의 분포가 각각 $N(n\mu, \sigma^2)$, Poisson($n\lambda$)가 됨을 알 수 있다.

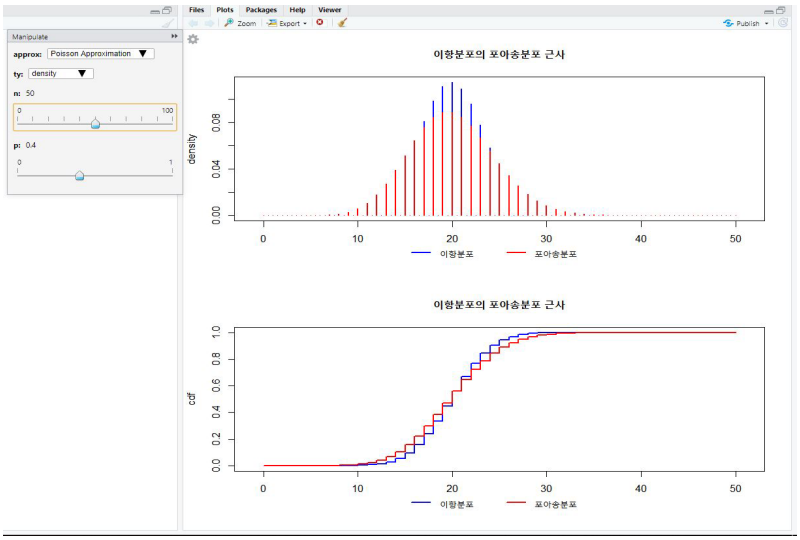
3.2. 이항분포의 포아송 근사

이항분포는 평균이 일정할 때 표본수가 커지면서 포아송 분포로 수렴한다. 또한 이항분포는 표본수가 커지면서 정규분포로 수렴한다. R 프로그램을 통해 표본수가 커지면서 이항분포와 포아송분포 또는 정규분포의 개형을 확률밀도함수 또는 누적분포함수를 비교할 수 있다. 이항분포 $B(n, p)$의 n과 p를 변경하면서 확률분포의 변화를 시각적으로 볼 수 있도록 manipulate 패키지를 이용하여 R 프로그램을 작성했는데 Figure 3.3과 같다.

Figure 3.3은 표본수 $n = 50$일 때 이항분포의 포아송근사를 보여준다. Figure 3.3(a)와 (b)는 각각



(a) $p = 0.05, n = 50$



(b) $p = 0.4, n = 50$

Figure 3.3. Poisson approximation to binomial distribution using R.

$p = 0.05$, $p = 0.4$일 때 이항분포와 포아송분포를 확률밀도함수와 누적분포함수를 같이 그려서 비교한 것이다. 메뉴에서 p가 매우 작을 때 n을 크게 하면 이항분포는 포아송분포에 근접하지만 p가 크면 이항분포는 포아송분포에 근접하지 않음을 볼 수 있다. p가 작을 때 n이 커지면서 일정한 결과를 보이는 것은 이항분포가 수학적으로 포아송분포로 근사하는 것을 증명할 수 있다는 것을 의미한다.

이항분포 $B(n, p)$의 포아송분포 Poisson($\lambda = np$)의 근사는 이항분포 적률생성함수의 극한이 포아송분포 적률생성함수와 같음을 wxMaxima를 이용하여 증명할 수 있는데 Figure 3.4와 같다. Figure

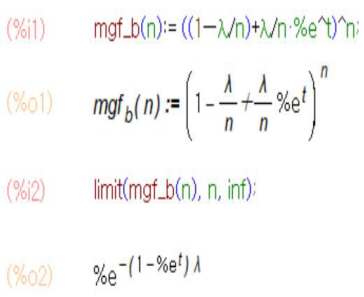


Figure 3.4. Poisson approximation to binomial distribution using wxMaxima.

3.4를 보면 이항분포의 적률생성함수에서 p를 λ/n로 변경하여 적률생성함수를 $\text{mgf}_b(n) = (1 - \lambda/n + \lambda/ne^t)^n$로 지정하고 $\lim_{n\to\infty} \text{mgf}_b(n)$을 구하면 포아송분포의 적률생성함수인 $e^{-(1-e^t)\lambda}$가 구해진다. 이를 통해 평균이 일정할 때 표본수가 커지면 이항분포가 포아송분포로 근사되는 것을 수학적으로 확인할 수 있다.

3.3. 최대가능도추정량의 도출

확률변수가 식 (3.1)의 지수분포 또는 식 (3.2)의 포아송분포를 따르는 확률표본일 때 λ의 최대가능도추정량(maximum likelihood estimator)을 구하는 것은 이론통계학의 대표적 문제이다.

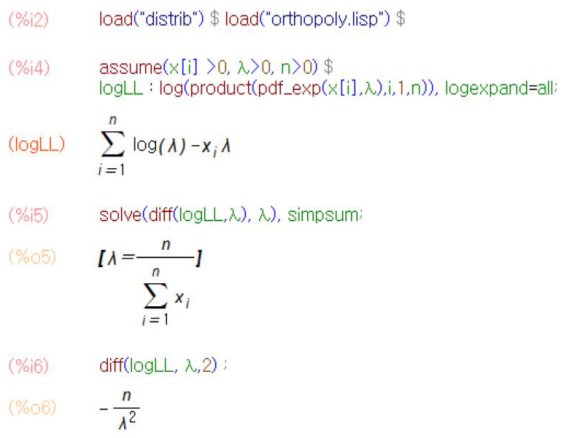
$$f(x|\lambda) = \lambda e^{-\lambda x}, \tag{3.1}$$

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{3.2}$$

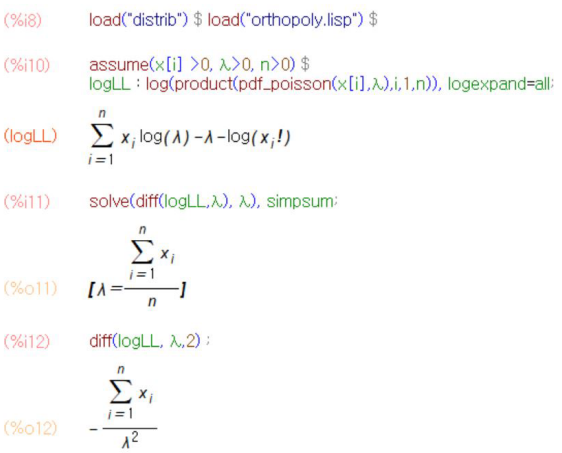
일반적 풀이는 Lee 등 (2019)에 보듯이 로그 가능도를 구하고 이를 1차 미분한 후 0으로 두고 최대가능도추정량을 구한 후 로그가능도를 2차 미분하여 도출한 최대가능도추정량이 최대가능도임을 확인하는 것이다.

이에 대한 wxMaxima 프로그램 수행 결과는 Figure 3.5(a)와 같다. 이를 보면 `load` 명령어로 관련된 패키지를 불러온다. `(%i2)`에서 `load("distrib") $ load("orthopoly.lisp")$`는 분포관련 패키지를 불러오고 것이고, `assume(x[i]>0,λ>0, n>0)$`는 관련 변수의 부호를 가정한 것이다. `logLL : log (product(pdf_exp(x[i],λ),i,1,n)), logexpand=all;`는 지수분포 확률밀도함수를 이용해서 로그가능도 함수를 구하고 이를 `logLL`로 지정한 것이다. `solve(diff(logLL,λ),λ ), simpsum;`는 로그가능도함수를 λ에 대해 미분한 후 0으로 두고 최대가능도추정량을 구한 것이다. `diff(logLL, λ, 2)`는 로그가능도 함수를 2번 미분한 것인데 그 결과를 보면 음수이다. 따라서 앞서 구한 표본평균의 역수가 λ의 최대가능도추정량임을 알 수 있다. 포아송분포 λ의 최대가능도추정량은 분포를 `pdf_exp` 대신 `pdf_poisson`로 바꾼 후 동일한 방식으로 구할 수 있는데, wxMaxima 프로그램 수행 결과는 Figure 3.5(b)와 같다.

Figure 3.6은 지수분포와 포아송분포에서 최대가능도추정량을 수치 해석적 방법으로 구하고, n과 lambda(λ)를 달리하여 최대가능도추정량과 가능도의 움직임을 그래프로 살펴볼 수 있는 그래프이다. Figure 3.6(a), (b)의 상단 그래프는 히스토그램과 실제 확률밀도(질량)함수가 나타나 있고, 하단 그래프에는 가능도함수와 구한 최대가능도추정량 값과 참값이 나타나 있다.



(a) Exponential Distribution



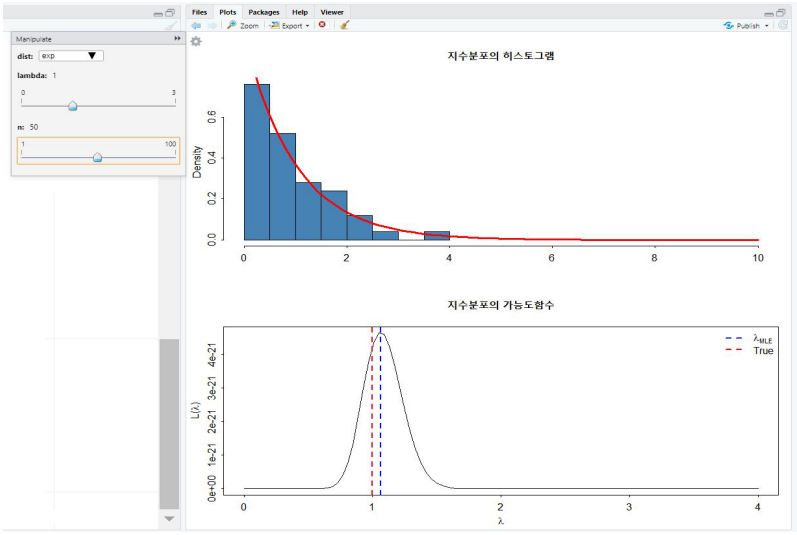
(b) Poisson Distribution

Figure 3.5. Maximum likelihood estimation using wxMaxima.

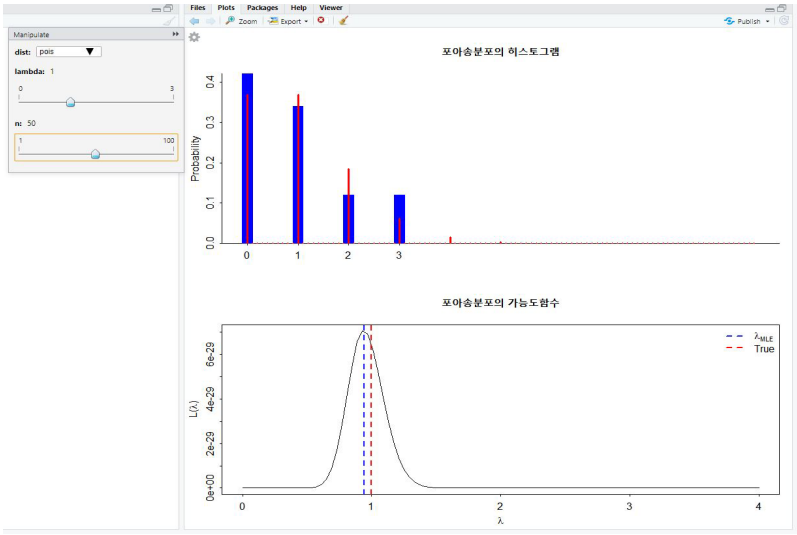
이와 같이 최대가능도추정량을 wxMaxima를 이용하여 수학적으로 도출해보고, R을 이용하여 시뮬레이션으로 시각적으로 다시 볼 수 있다. 학생들이 직접 wxMaxima와 R 코드를 작성하여 문제를 동시에 접근하여 해결한다면 학생들이 이론통계학 교육으로부터 통계적 원리를 보다 포괄적으로 이해할 수 있을 것으로 판단된다.

4. 맺음말

통계학과의 교육은 수학을 기반으로 통계학 원리를 증명하는 이론통계학 (수리통계학)이 통계학 교육의 중심이 진행되어 왔다. 이는 20세기 통계학의 원리가 수학을 기반으로 발전한 데 기인한다 (Efron과 Hastie, 2016). 따라서 이론통계학의 강의는 2개 학기에 걸쳐 수학적 증명 및 연산을 중심으로 한 칠판



(a) Exponential Distribution



(b) Poisson Distribution

Figure 3.6. Maximum likelihood estimation using R.

강의로 주로 진행되어 왔으며, 학생들은 통계학 원리를 수학 기반 과제 또는 시험을 통해 학습하여 왔다. 학생들이 전통적 방식의 이론통계학 교육으로 이론통계학을 학습할 때 대학수학(미적분)에 대한 지식이 부족한 경우 이론통계학을 학습하기 어렵다. 또한 수학을 이용하여 문제를 증명했더라도 통계학의 핵심 원리를 이해하지 못하는 경우가 많다. 이론통계학 교육에서 수학이란 장벽을 극복하는 한 방법으로는 일반적인 수식을 다룰 수 있는 수학소프트웨어 CAS의 도움을 받아 반복적으로 관련 문제를 푸는 것이다. 이를 통해 이론통계학 교육을 어렵게 만드는 수학의 두려움에서 벗어나 통계학의 원리 그 자체를 대

면할 수 있다. 이 논문에서는 수학소프트웨어 wxMaxima와 R을 이용한 시뮬레이션을 통해 통계학 핵심 원리를 반복해서 수학으로 풀어보고, 시각적으로 살펴볼 수 있는 방법을 확률변수 합의 분포, 이항분포의 포아송(정규) 분포 근사, 최대가능도 추정이라는 3개의 예를 통해 살펴보았다.

학생들이 이론통계학을 보다 심도 깊게 이해하기 위해서는 시뮬레이션, 수학을 이용한 문제 풀이에 더해서 역사와 철학을 고려한 인문학적 접근이 필요하다. 예를 들면 고셋(Gosset)이 t분포를 발견하고 이를 피셔(R.A. Fisher)가 증명하면서 논의되었던 것들을 그 시대(20세기 초 영국) 상황과 같이 이해한 후 그 t분포가 이후 어떻게 활용되었는지를 글 또는 영상으로 정리하는 과제를 학생들에게 주는 것이다. 학생들이 통계학 원리를 인문학적 검토, R을 이용한 시뮬레이션, wxMaxima를 이용한 수학적 증명을 결합하여 학습한다면 이론통계학 교육이 수학을 이용한 문제 풀이에 매몰되지 않고 통계학 원리를 보다 개념적으로 이해할 수 있는 교육으로 변할 수 있을 것으로 판단된다.

부록: R 코드

A.1. 합과 평균의 분포

```
library(distrEx)
library(manipulate)
# 합과 평균의 분포 생성
dist1 = function(distr,n, type1, type2){
 Y = X_1 = distr
 for (i in 1:(n-1)) Y = Y + distr
  if (type2=="mean") {X = Y/n
    } else {X = Y}
  ty  = ifelse(type1=="pdf", "d", "p"); norm1 = ifelse(type1=="pdf", dnorm, pnorm)
 sty1 = ifelse(type1=="pdf", expression(paste(X[i], " 의 확률밀도함수 : %C(%P)")),
                          expression(paste(X[i], " 의 누적분포함수 : %C(%P)")))
if(type2=="sum"){
 sty2 = ifelse(
   type1=="pdf", expression(paste(sum(X[i], i==1, n), " 의 확률밀도함수 : %C(%P)")),
                 expression(paste(sum(X[i], i==1, n), " 의 누적분포함수 : %C(%P)")))
  } else{
 sty2 = ifelse(type1=="pdf", expression(paste(bar(X), " 의 확률밀도함수 : %C(%P)")),
       expression(paste(bar(X), " 의 누적분포함수 : %C(%P)")))
}
 sty3 = ifelse(type2=="sum","개 %C(%P) 확률변수 합의 분포",
              "개 %C(%P) 확률변수 평균의 분포")
 max1 = max(distr::q(X_1)(0.99), distr::q(X)(0.99))
 min1 = min(distr::q(X_1)(0.01), distr::q(X)(0.01))
par(mfrow=c(2,1))
plot(X_1,mfColRow = FALSE,to.draw.arg=c(ty), ylab="", xlab="", xlim=c(min1, max1),
     cex.points = 1.0, main = paste(n, sty3),  inner = list(sty1))
 legend("bottom", c(paste0("기댓값 :", format(distrEx::E(X_1), nsmall=2)),
```

```
                  paste0("분산 :",  format(distrEx::var(X_1)), nsmall=2),
                  paste0("왜도 :",  format(distrEx::skewness(X_1), nsmall=2)),
                  paste0("첨도 :",  format(distrEx::kurtosis(X_1), nsmall=2))),
         col="white", bty="n", xpd = TRUE, horiz = TRUE, inset = c(-0.4,  -0.4))
plot(X,  mfColRow = FALSE,to.draw.arg=c(ty), ylab="", xlab="", main="",
     cex.points = 1.0,  xlim=c(min1, max1), inner = list(sty2))
 legend("bottom", c(paste0("기댓값 :", format(distrEx::E(X), nsmall=2)),
                  paste0("분산 :",  format(distrEx::var(X)), nsmall=2),
                  paste0("왜도 :",  format(distrEx::skewness(X), nsmall=2)),
                  paste0("첨도 :",  format(distrEx::kurtosis(X), nsmall=2))),
         col="white", bty="n", xpd = TRUE, horiz = TRUE, inset = c(-0.4,  -0.4))
}
# GUI 확면 생성
manipulate(
 stat = picker("sum", "mean"),
 distr = picker("Norm", "Exp", "Chisq", "Binom", "Pois"),
 n = slider(2, 100, step=1, initial=16, label="sample size"),
 type = picker("pdf", "cdf"),
if (distr=="Norm") {dist1(Norm(1,1),n, type, stat)
 } else if(distr=="Exp") {dist1(Exp(1),n, type, stat)
 } else if(distr=="Chisq") {dist1(Chisq(1),n, type, stat)
 } else if(distr=="Binom") {dist1(Binom(1,0.5),n, type, stat)
 } else {dist1(Pois(1),n, type, stat)
 })
```

A.2. 이항분포의 포아송근사와 정규근사

```
library(manipulate)
options(warn = -1)
# 이항분포의 포아송 근사
a_dist = function(ty,n,p){
  ty1 = ifelse(ty=="density", "h", "s")
  dist1 = ifelse(ty=="density", dbinom, pbinom)
  dist2 = ifelse(ty=="density", dpois, ppois)
  x = 0:n
  plot(x, dist1(x,n, p), type=ty1, ylab=ty, xlab="", xlim=c(-1, (n+1)),
       lwd=2, col=4, main="이항분포의 포아송분포 근사")
  curve(dist2(x,n*p), 0, n, lwd=2, col=2, type=ty1, add=TRUE)
  legend("bottom", c("이항분포", "포아송분포"), col=c(4,2), lwd=2, bty="n", xpd = TRUE,
         horiz = TRUE, inset = c(-0.3,  -0.3))
}
# 이항분포의 정규 근사
```

```
n_dist = function(ty,n,p){
  ty1 = ifelse(ty=="density", "h", "s")
  dist1 = ifelse(ty=="density", dbinom, pbinom)
  dist3 = ifelse(ty=="density", dnorm, pnorm)
  x = 0:n
  plot(x, dist1(x,n, p), type=ty1, ylab=ty, xlab="", xlim=c(-1, (n+1)),
       lwd=2, col=4, main="이항분포의 정규분포 근사")
  curve(dist3(x,n*p, sqrt(n*p*(1-p))), 0, n, lwd=2, col=2, type="l", add=TRUE)
  legend("bottom", c("이항분포", "정규분포"), col=c(4,2), lwd=2, bty="n",
         xpd = TRUE, horiz = TRUE, inset = c(-0.3,  -0.3))

}
# GUI 확면 생성
par(mfrow=c(2,1))
manipulate(
  approx = picker("Poisson Approximation", "Normal Approximation","both"),
  ty = picker("density","cdf"),
  n = slider(0, 100, initial=10,  step=10),
  p = slider(0,   1, initial=0.05, step=0.01),
  if (approx=="Poisson Approximation"){
    a_dist("density",n,p)
    a_dist("cdf",n,p)
  } else if (approx=="Normal Approximation") {
    n_dist("density",n,p)
    n_dist("cdf",n,p)
  } else {
    a_dist(ty,n,p)
    n_dist(ty,n,p)
  })
```

A.3. 최대가능도추정

```
library(stats)
library(manipulate)
library(arm)
# 확률분포 지정
dist_mle = function(n, lambda1, dist){
if (dist=="exp") {
 y <- rexp(n, rate=lambda1); main1="지수분포"
 nLL <- function(lambda) -sum(dexp(y, lambda, log = TRUE))
} else {
 y <- rpois(n, lambda=lambda1); main1="포아송분포"
```

```
 nLL <- function(lambda) -sum(dpois(y, lambda, log = TRUE))
}
# 최대가능도추정
fit <- mle(nLL, method = "Brent", start = list(lambda = 1),
            lower=0, upper=20, nobs=NROW(y))
print(summary(fit))
mle_y <- fit@coef
lambda = seq(0, (lambda1+3), length=100)
t1 =paste0(main1, "의 히스토그램"); t2 =paste0(main1, "의 가능도함수")
# 가능도함수 생성
par(mfrow=c(2,1))
if (dist == "exp") {
  LL  <- lambda^n*exp(-lambda * sum(y))
  hist(y, main=t1, freq=FALSE, xlab="", col="steelblue", xlim=c(0,10),
       border=TRUE, nclass=10)
  curve(dexp(x,lambda1),  0, 10, lwd=3, col=2, type="l", add=TRUE)
} else {
  LL  <- lambda^sum(y)*exp(-lambda*n)/prod(factorial(y))
  discrete.histogram(y, main=t1, freq=FALSE, xlab="", col="steelblue",
    xlim=c(-0.5,10.5))
    curve(dpois(x,lambda1), 0, 10, lwd=3, col=2, type="h", add=TRUE)
}
plot(lambda, LL, type="l",  xlab=expression(lambda),
     ylab=expression(paste("L(", lambda, ")")), main=t2)
 abline(v=mle_y,  lty=2, lwd=2, col=2)
 abline(v=lambda1,lty=2, lwd=2, col=3)
 legend("topright", c(expression(lambda[MLE]), "True"), lty=c(2,2), col=c(4,2),
    lwd=2,  bty="n") }
# GUI 화면 생성
manipulate(
 dist = picker("exp", "pois"),
 lambda = slider(0, 3, step=0.05, initial=1),
 n = slider(1, 100, step=1, initial=30),
 dist_mle(n,lambda, dist)
)
```

References

Baglivo, J. (1995). Computer Algebra Systems: maple and mathematica, *The American Statistician*, **49**, 235–249.

Berger, R. L. (1998). Using computer algebra systems to teach graduate mathematical statistics: Potential and pitfalls, *Statistical Education - Expanding the Network: Proceedings of the Fifth International*

Conference on Teaching of Statistics (eds. Pereira-Mendoza, L., et al.). International Statistical Institute, Voorburg, Netherlands. Volume 1, 189–195.

Carver R., Everson, M., Gabrosek, *et al.* (2016). *Guidelines for assessment and instruction in statistics education: College report*, ASA GAISE College working group.

Doi, J., Potter, G., Wong, J., Alcaraz, I., and Chi, P. (2016). Web Application Teaching Tools for Statistics Using R and Shiny, *Technology Innovations in Statistics Education*, **9**. Available from: https://escholarship.org/uc/item/00d4q8cp

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference Algorithms, Evidence, and Data Science*, Cambridge University Press.

Green, J. L. and Blankenship, E. E. (2015). Fostering conceptual understanding in mathematical statistics, *The American Statistician*, **69**, 315–325.

Horton, N. J., Brown, E. R., and Qian, L. (2004). Use of R as a toolbox for mathematical statistics exploration, *The American Statistician*, **58**, 343–357.

Hunter, D. R. (2005). Teaching computing in statistical theory courses, *The American Statistician*, **59**, 327–333.

Jang, D. H. (2009). Application of R for inferential statistics in the elementary Statistics Course, *Journal of Applied Statistics*, **22**, 893–910.

Kang, K. and Park, J. (2015). *Mathematical Statistics-Practice with R*, Freedom Academy, Seoul.

Kim, W. C. (2011). *Mathematical Statistics*, Yougjisa, Seoul.

Kumar, A. and Kumaresan, S. (2008). Use of Mathematical software for teaching and learning mathematics, *Proceedings of 11th International Congress on Mathematics Education*, Mexico.

Lee, G. H., Kim, H., Kim, J., Park, J., and Lee, J. (2019). *Concepts and Controversies of Statistics*, KNOU Press, Seoul.

Lee, S. G. and Park, K. E. (2015). Improving computational thinking abilities Through the teaching of mathematics with Sage, *Communications of Mathematical Education*, **29**, 19–33.

Leydold, J. and Petry, M. (2011). *Introduction to Maxima for Economics*, Institute for Statistics and Mathematics, WUWien.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics, *International Statistical Review*, **65**, 123-165.

Rose, C. and Smith, M. D. (2000). Symbolic maximum likelihood estimation with Mathematica, *Journal of the Royal Statistical Society. Series D (The Statistician)*, **49**, 229–240.

Rossman, A. and Chance, B. (1999). Teaching the reasoning of statistical inference: A "Top Ten" List, *The College Mathematics Journal*, **30**, 297–305.

이론통계학 교육에서 수학 소프트웨어의 활용

이긍희[a,1]

[a]한국방송통신대학교 정보통계학과

(2019년 5월 17일 접수, 2019년 6월 6일 수정, 2019년 6월 13일 채택)

요 약

이론통계학은 통계학의 원리를 수학을 이용하여 배우는 교과목이다. 학생들이 수학을 충분히 알지 못하는 경우 이론통계학 교육을 통해 통계학의 원리를 이해하는 데에는 제약이 있다. 이론통계학 교육을 통해 통계학의 원리에 대한 이해를 높이기 위해 수학적 문제풀이 외에 R 프로그램을 이용한 통계 시뮬레이션이 보조적으로 도입되어 왔지만 수학을 이용한 문제풀이를 대신하지는 못하고 있다. 이 논문에서는 wxMaxima, Wolfram Alpha 등 기호 수학 연산이 가능한 수학 소프트웨어 CAS를 소개하고, 이를 이용하여 이론통계학 교육에 걸림돌이 되는 수학의 어려움에서 벗어나 통계학의 원리 자체를 학습할 수 있는 방안을 모색하였다.

주요용어: 수리통계학, 통계적 추론, CAS, wxMaxima, Wolfram Alpha, R.

이 논문은 2017년 한국방송통신대학교 3/4분기 학술연구비의 재정지원을 받아 작성된 것임.

[1](03087) 서울특별시 종로구 대학로 86, 한국방송통신대학교 정보통계학과. E-mail: geunghee@knou.ac.kr