대한수학교육학회지 <학교수학> 제 12 권 제 3 호
Journal of Korea Society of Educational Studies in Mathematics
School Mathematics Vol.12, No.3, 273-299. Sep 2010

통계적 추론에서의 표집분포 개념 지도를 위한 시뮬레이션 소프트웨어 설계 및 구현

이 영 하* · 이 은 호**

본 논문의 목적은 고등학교 수준의 학생들이 표집분포의 개념을 학습할 수 있도록 '표집분포 시뮬레이션 (Sampling Distributions Simulation)'을 설계하고 구현하는 것이다. '표집분포 시뮬레이션'은 다음과 같이 4차시로 구성되어 있다.

1차시 – 신뢰도와 신뢰구간의 의미 학습하기
2차시 – 표집분포의 의미 학습하기
3차시 – 중심극한정리의 의미 학습하기
4차시 – 이항분포의 정규근사 학습하기

본 연구를 통하여 표집분포의 중요성에 대한 학생들이 인식이 달라지고 이해가 증진되기를 기대한다. 또 본 연구의 결과로 제공되는 프로그램 '표집분포의 시뮬레이션' 수업을 통해 통계적 추론 능력이 향상되고, 아울러 통계적 추론 속에서 표집분포의 역할이 충분히 이해되기를 기대한다.

I. 서론

표집분포[1](sampling distribution)란 무엇이고, 왜 중요한가? 표본통계량(sample statistic)의 확률분포(probability distribution)를 표집분포라고 하는데, 고등학교 통계단원의 통계적 추론(statistical inference)이 관련된 설명에서 표집분포를 소개하지 않은 채로 설명을 하면 어떤 설명을 해도 추론의 논리성을 이해하는데 아무 소용이 없다. 즉 통계적 추론의 과정을 이해하는데 있어서 표집분포 개념이 없이는 설명이 불가능할 정도로 표집분포는 매우 중요한 개념이라는 것이다.

1. 표집분포의 개념학습 문제

한편, 이 개념은 개념의 표현 관행상 매우 이해하기 어려운 개념이라는 것과, 더욱이 실용적인 표본 통계량의 경우에, 고교 수준의 수학적 계산으로는 도출이 거의 불가능할 정도라는 것이다. Cobb & Moore(1997)도 표집분포 개념 이해의 어려움이 통계적 추론을 이해하는데 있어 장벽이 된다고 주장하였다.

개념 표현 관행 상 표집분포의 이해가 어렵다는 것을 이해하기 위해 다음과 같은 표현을 생각해보자. "어느 모집단에서 얻은 크기 100인 임의표본(random sample)의 표본평균(sample mean)은 근사적으로 평균이 50, 표준편차가 10인 정규분포를 한다." 이 표현이 뜻하는 통계 조사 상의 의미는 "모집단에서 크기 100인 표본을 반복해서 추출하여 모든 가능한 추출의 경우(모집단의 변량의 개수가 N개[2])라면, 비복원 추출일 때 총

* 이화여대 수학교육과 (youngha@ewha.ac.kr)
** 이화여대 교육대학원 (lovejean98@hanmail.net)
1) 표본분포함수(sample distribution function, empirical distribution function)와는 다름

NC_{100} 가지)마다 구한 평균들의 도수분포에 근거한 확률분포가 근사적으로 평균이 50, 분산이 100인 정규분포"라는 뜻이다. 즉 여기서의 표본 통계량은 표본평균이고, 표집분포는 정규분포인 것이다.

그런데 통계조사와 관련하여, 전자의 이 표현이 후자의 이런 의미를 갖는다는 것과 실제 통계조사와는 아무 관련이 있어 보이지 않는다는 것이 학생들이 겪는 인식론적 오류(epistemological fallacy)의 핵심이다. 가령 어느 학교 학생들(모집단) 중에 100명(표본의 크기)의 키를 조사할 때, 학생들은 어떤 경험으로도 단 한번 100명의 키를 조사할 뿐, 또 다른 100명에 대해 이를 반복하지는 않으며, 따라서 표본평균은 그렇게 엄청나게 많은(NC_{100}개) 값이 구해지는 것이 아니라 단 한 개 구해질 뿐인데, 그 표본 평균이라는 숫자 한 개로는 어떤 의미로도 분포라는 것을 도출할 수가 없기 때문에 후자의 의미를 이해할 수 없게 되는 것이다. 그럼에도 불구하고 표집분포의 의미를 실제의 통계조사 상황과 연결 지어 이해하려면, 그 통계적 추론에 관련된 통계량 표집분포의 역할을 이해하여야 하기 때문인데, 평균값 계산이나 그래프 그리기 등 지엽적인 것에 열중하고 있는 학생들에게 표집분포를 포함한 전체적인 것을 보아야 하는 이 문제는 어렵게만 느껴질 것이 당연해 보인다.

둘째 수학 계산적 측면에서도, 가령 1부터 100만까지의 자연수가 한 개씩 적혀 있는 어느 모집단에서 크기 100인 표본의 표본 평균의 표집분포를 수학적으로 구하라하면, 우선 엄청나게 많은 경우에 대해 각각 표본 평균을 먼저 구해놓고 그들의 분포를 생각해야 한다는 것도 상상하기조차 싫은 일이지만, 더욱이 어떤 표본들은 100개씩의 구성하는 값들의 차이에 관계없이 평균값은 같을 수 있고, 그러면 같은 평균값이 각각 총 몇 개씩인가에 따라 표본평균의 도수분포가 얻어질 것이며, 그래야 이와 관련된 확률분포도 구할 수 있을 것이라는 생각을 한다면, 대부분의 사람들은 고지식한 방법으로는 구할 수 없을 것이라는 생각을 할 것이다. 즉 이 문제는 계산적으로도 난감해 보이게 되는 데, 일부 실생활과 관련된 특별한 경우에 대하여 수학적 해법이 잘 알려져 있으나, 이를 고등학교에서 도입하기에는 미리 준비할 내용이 너무 많고, 그 준비내용 역시 학생들이 잘 이해할 수 있으리라는 보장도 없다.

여기서 문제 해결 방법으로 떠오르는 것이 컴퓨터이다. 왜냐하면 이 문제의 어려움의 핵심이 생각해야할 경우들의 가짓수가 사람으로서는 다루기에 너무 방대하다는 것, 그러나 컴퓨터로서는 가장 간단한 문제이기 때문이다.

통계교육이라는 관점에서 이런 차이는 매우 특별한 의미를 갖는다. 흔히 수학적 방법은 그 해결과정이 매우 기술적(technical)인 경우가 많다. 여기서 기술적이라 함은 "풀이(문제해결) 과정이 찾고자하는 답을 향해 이동해 가고 있다는 느낌이 전혀 들지 않은 채, 상식적으로는 이해할 수 없는 과정을 한참 따라가다 보면, 갑자기 답에 도달했다는 주장만 듣게 되는" 과정이라는 의미이다. 가령 방정식의 근을 구하려면 이 값 저 값 등식이 성립하도록 해 주는 미지수의 값을 찾아 헤매는 느낌이 드는 과정이어야 상식적인 풀이 과정인데, 그와는 전혀 동떨어진 이상한(등식의 성질을 이용한) 계산을 하다가 갑자기 답을 얻었다고 하여, "식에 대입해보니 등식이 정말로 성립하더라!"와 같은 경우를 기술적 해법의 대표적인 예로 들 수 있을 것이다. 표집분포를 수학적으로 구하는 것도, 마찬가지로, 이렇게 기술적이기 때문인데, 컴퓨

2) N=200정도라면 전수조사를 할 것이다. 적어도 1000이상의 매우 큰 수로 본다.

터의 모의실험(simulation)을 이용하는 경우에는 수학적 기술을 발휘하는 것이 아니라 고지식하게 표집분포를 그 의미대로(같은 크기의 표본들을 반복적으로 매우 많은 횟수 추출해서) 구해나간다는 점에서 교육적으로 학생들에게 가장 이해가 용이한 방법이라고 할 수 있기 때문이다.

통계적 방법에서 가장 핵심이 되는 것은 통계적 추론을 통한 합리적 의사결정이다. 이에 대해 우정호(2000)는 "합리적 의사결정을 위한 판단 도구로써 통계적 방법이 놀라울 정도로 광범위하게 사용되고 있으며, 학문 연구와 기술개발에도 통계적 방법은 매우 널리 이용되고 있어 통계적 소양이 현대 사회생활을 영위하는 필수적 소양이 되었다."고 적고 있다.

우리나라 제7차 교육과정의 '확률과 통계'는 학생들이 이러한 정보화 시대에 필요한 자료 처리 능력과 통계적 추론 능력을 신장시키고 여러 가지 확률 통계적 사회 현상 및 자연 현상을 이해하고 해석하는 능력과 태도를 기르는 것을 목표로 하고 있다(강행고, 1998). 이는 학생들이 실생활에서 접하는 자료를 효율적으로 조사, 정리, 분석해 봄으로써 유용한 정보를 얻는데 효과적인 도구가 통계적 방법임을 알 수 있게 지도하도록 요구하고 있는 것이다. 이렇게 많은 학자들은 통계적 추론을 반드시 수학교육과정에 포함시킬 것을 주장하고 있다.

한편 통계적 추론은 가르치되, 앞서 언급했던 여러 어려움을 회피하는 한 가지 교육학적 대안으로서, 현행 교육과정에서처럼, 교육과정에서 표집분포에 관한 논의를 피해 가는 방법을 생각할 수 있다. 그러나 그것은 결코 대안이 될 수 없다. 즉, 표집분포를 우회하여 통계적 추론을 가르칠 수 있는 방법은 없다. 그 이유는 통계적 추론이 표본 자료 값들에 관한 추론처럼 보일지라도, 사실상 본질적인 것은 표집분포 자체에 관한 추론이기 때문이다.

이것은 서울시청(표본 자료의 값)의 위치를 설명하되, 서울시 지도(표집분포) 또는 그 개념은 사용하지 않고 설명하려는 것과 같다. 시청의 위치를 설명하므로 지도는 관계없는 듯하지만, 시청은 지도 속에서 파악될 때 비로소 그 상대적 위치 개념이 생기기 때문이다. 즉 표본 자료에서 계산된 숫자로 된 한 개의 평균값이 큰지, 작은지는 그 표본평균의 표집분포와 비교했을 때 비로소 답할 수 있기 때문이다.

우리나라 통계교육과정에서는 표집분포에 대한 개념을 도입하지 않고, 단순히 두 개의 분포, 즉 이항분포와 정규분포에 대한 소개에 치중하게 됨으로써, 결과적으로 표집분포 개념 자체를 간단하게 설명하거나 아예 생략하고 마는 결과를 초래하고 있다(윤현진 외, 2009).

그 대안으로, 연구자는 중심극한의 정리를 통해 표본평균의 표집분포를 지도할 것을 주장하며, 그리하여 고등학교 모평균의 신뢰구간의 지도에서 "모평균을 모르는데 어떻게 모집단의 분포가 정규분포인지를 아느냐?"는 학생들의 의문을 일부라도 해소시켜 줄 수 있기를 바라는 것이다.

2. 컴퓨터를 이용한 해결 방안

컴퓨터를 이용하여 수학적 개념 지도의 어려움을 경감할 수 있는 방안에 대한 광범위한 연구가 진행되는 등, 수학 교수-학습 과정에서 제기되는 여러 가지 어려움을 극복하기 위한 대안으로 많은 사람들이 컴퓨터에 주목하고 있다.(Freudenthal(1981), Tall, D.(1991), 우정호(2000), NCTM(2000))

통계교육에서 컴퓨터 활용을 주장하는 연구도 많은데, (나귀수(2000), 강문봉 외(2003), 장형도(2006)) 특히, 통계교육 등에서 시뮬레이션

방법의 활용을 촉구하는 연구와 주장도 많다. (NCTM(1989), 이종영(1999), 우정호(2000), 손미선(2003))

본 연구는 이런 여러 주장에 공감하여, 표집분포 지도에서 이를 구체적으로 구현해 낼 방안을 찾고자하는 것이다.

우리나라는 고등학교까지의 통계교육과정에서 실질적 의미의 표집분포 개념을 다루고 있지 않기 때문에 학생들에게 표집분포의 개념을 이해시키기 위한 연구 또한 거의 이루어지지 않고 있다. 더불어 표집분포 개념 이해를 위한 어떠한 시도(모의실험법 조차)도 전혀 개발되어 있지 않다.

본 연구에서는 이러한 수학적 절차를 대신하여, 무한개의 표본평균을 얻는 대신 시뮬레이션을 이용하여 컴퓨터로 하여금 매우 많은 수의 표본평균(본 논문에서는 2,000개)을 구하고, 그것으로 밀도도수히스토그램을 그리도록[3] 하여, 학생들이 표본평균의 확률밀도함수를 근사적으로 느끼도록 함으로써 표집분포에 대한 직관적 이해를 갖게 하고자 한다. 따라서 본 연구의 목적은 고등학교 수준의 학생들이 표집분포 개념을 쉽게 이해할 수 있도록 사용법이 쉽고 편리하면서, SPSS나 SAS 같은 값비싼 소프트웨어에 의지하지 않는 컴퓨터 모의실험 도구인 '표집분포 시뮬레이션 (Sampling Distributions Simulation)'을 개발하여 shareware로 공개함으로써 고교에서의 표집분포 지도 관련 학교 통계교육을 지원하고자 하는 것이다.

II. 연구 세부사항과 이론적 배경

1. 통계적 추론과 표집분포

본 연구가 궁극적으로 추구하는 "표집분포 시뮬레이션" 개발만을 생각한다면, 표집분포의 통계적 추론과의 관련성은 반드시 이론적으로 검토할 필요는 없다. 표집분포의 뜻에 맞도록 컴퓨터를 조작하는 명령문만을 나열하면 충분하기 때문이다.

그런데 기왕 학교 교육용으로 제작되는 시뮬레이션 프로그램이라면, 통계학에 나타난 표집분포의 실용적 의미와 활용을 문제의 배경 상황으로 하여 프로그램을 제작하는 것이 더 적절하다고 할 수 있을 것이므로, 본 연구에서는 그런 상황에 맞추어 프로그램을 개발하려고 한다.

한편 본 연구 결과의 시뮬레이션 프로그램을 이용하여 표집분포 수업을 하려는 교사에게 있어서는 이 프로그램의 사용법 차원에서 표집분포의 내용 구현에 대한 상세한 이해가 필요한데, 특히 개발된 프로그램이 앞서 말한 바와 같은 것이라면, 결국 통계적 추론에서 표집분포의 역할에 대한 이론적 배경을 확실히 해 두는 것이 필요하다.

통계적 추론과 표집분포와의 상호관계는 앞서 서울시 지도의 비유를 통해 매우 간단하게 언급되었지만, 대표적인 두 가지 통계적 추론, 즉 가설검정과, 신뢰구간과 관련하여, 통계적 문제에서 표집분포가 어떤 기능과 역할을 갖는지에 대해 알기 위해서는, 상세한 수학적 절차는 생략하고, 추론의 흐름만을 전체적 윤곽을 통해 조망할 때 더욱 분명해 지리라고 본다.

우선 모평균의 신뢰구간에 대하여는 표본 자료 $X_1, X_2, X_3, \ldots, X_n$를 토대로 구하는 통계량 S

3) 이것은 도수분포와 확률분포 간의 연결이 잘 이루어져야 이해가 가능한 설명이다. 밀도도수히스토그램은 확률밀도함수와 유사한 것이나 이에 관한 설명이 현행 교육과정에는 없다. 이를 위한 보완 내용은 뒤의 분포개념 부분에서 보충 설명하기로 한다.

를 대개 $S(X_1, X_2, \ldots, X_n) = \frac{\Sigma X_i}{n} = \overline{X}$, 즉 표본평균을 사용한다. 표본 평균 S의 표집분포는, 표본크기 n의 값이 크고, 모평균이 m, 모분산이 σ^2이라면, 중심극한의 정리에 따라 평균이 m이고 분산이 $\frac{\sigma^2}{n}$에 가까운 정규분포를 한다.[4] 그러면 다시 $\frac{\sqrt{n}(\overline{X}-m)}{\sigma}$이 근사적으로 표준정규분포[5]를 한다고 할 수 있다. 표준정규분포 표를 이용하여 $P(-1.96 < \frac{\sqrt{n}(\overline{X}-m)}{\sigma} < 1.96) \approx 0.95$의 표현에서 확률기호 괄호안쪽 부분을 변형하여, $P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < m < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) \approx 0.95$에서, 모평균 m에 관한 95% 신뢰구간 $\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$을 얻는다. 즉 이 신뢰구간을 얻는데 있어서 가장 결정적인 것은 표본평균 S의 표집분포에 관한 진술, 즉 중심극한의 정리이다.

일반적으로 가설검정에서의 추론의 과정은 다음과 같다. ①먼저 검증하려는 가설 N과 A가 있는데, A를 주장하려는 가설, N을 그에 반대되는 가설이라 하자. ②검증을 위해 표본 자료 $X_1, X_2, X_3, \ldots, X_n$를 얻을 터인데, 이 자료로부터 위의 두 가설의 진위를 검증하는데 의미가 있어 보이는 통계량 $S = S(X_1, X_2, \ldots, X_n)$을 구하기로(대개 수리통계학적 이론에 의해 공식 S의 모양이 정해짐) 정한다. ③여기서 N이 참이라는 가정 하에서 S의 표집분포를 구해둔다. ④실제 자료수집이 이루어져서 $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$이라는 숫자로 된 자료 $x_1, x_2, \ldots, x_n$ 이 얻어지면 $s = S(x_1, x_2, \ldots, x_n)$이라는 통계가 이 자료로부터 계산 산출된다. ⑤산출된 통계치 s가 너무 크거나 작은지 등을 판단하기 위해 S의 표집분포(N이 참이라고 가정한 상태)에서 숫자 s의 위치를 살핀다. ⑥이 위치가 S의 표집분포 내에서 극단적[6]이라고 생각되면, N이 참인 상태에서는 숫자 s는 확률적으로 자주 발생할 것 같은 값이 아니라고 판단하게 된다. ⑦그리고 그럴듯하지 않은 값이 관찰된 것을 모순적 갈등상황으로 보아, 이 갈등이 N이 참이라고 가정한 데에서 비롯된(그런 타당해 보이지 않은 표집분포를 얻게 된) 것이라고 추론함으로써, 가설 N을 기각하고 A를 채택하게 되는 것이다. 즉 얻어진 통계치 s가 얼마나 그럴듯한가를 판단하기위해 N상태에서의 S의 표집분포를 s를 위한 지도처럼 사용하여, 이상한 점(가능성이 희박한 표본이 관찰되는 일)이 발견되면, 이 지도에 문제가 있다고 판단하는 형태인 셈이다. 결국 이 판단의 핵심에도 역시 "N상태에서의 표집분포"가 자리하고 있으며 이것을 구하는 것이 가설검정의 핵심이라고 할 수 있다.

이상의 과정에서 표집분포의 통계적 추론 관련 기능과 역할을 살펴보았는데, 다시 근본적인 문제, 즉 표집분포는 왜 그렇게 이해하기 어려운지를 살펴볼 필요가 있다. 이 문제는 본 연구의 프로그램 제작에 있어서 컴퓨터가 표현해 주어야 할 내용, 학생이 응답해야 할 내용 등 다소 사소해 보이는 것들을 정하는데 있어서 매우 중요하다. 즉 학습 후 학습자의 머릿속에 남게 될 학습자의 이해를 고려할 때, 프로그램 학습 중 학습자의 행동은 실제로는 매우 중요한 역할을 한다고 할 수 있기 때문이다.

---

4) 이 진술을 증명해야 하는데, 본 연구는 학생들이 컴퓨터를 이용하여 이것이 사실임을 느낄 수 있는 시뮬레이션 프로그램을 개발하려는 것이다.

5) 이 사실에 대한 증명이나 이해는 자연계 학생의 경우, 정규분포 확률밀도함수의 구간 a, b사이의 확률 구하기 문제에서, 치환적분을 통해 피적분함수가 표준정규분포 확률밀도함수로 바뀌는 것에 유의하도록 하여, 확률분포와 확률밀도함수 사이의 관계를 설명하면 어느 정도 수학적으로 이해가 가능하다.

6) S의 표집분포에서 S, s가 N, A와 관련된 확률 L(s)를 s의 관측유의수준(회귀도)라고 하는데 이 값이 유의수준(대개 5%)보다 작을 경우를 표집분포 내에서 s의 위치가 극단적이라고 판단함.

본 연구의 시뮬레이션 프로그램 제작에서 연구자는 시뮬레이션 프로그램 학습을 통해 학생 스스로가 무엇을 하고 있는지에 대한 경각심을 잃지 않도록 하는데 특별히 유의하였다. 계산이 복잡하다거나, 단순 반복이라거나 등의 이유로 프로그램이 컴퓨터에게만 일을 시키는 형태로 제작된 프로그램에서는, 자칫 학습자가 자신이 하고 있는 일의 중요한 단계 일부를 지나치기 쉽고, 그 결과 자칫 자신이 학습과 관련하여 무슨 일을 하고 있는지를 망각하게 될 수 있기 때문이다.

표집분포를 구하는 것은 통계적 추론과 관련하여 본질적으로 무엇을 구하는 행위이며, 어려움의 세부적 원인은 무엇인가?

우선 가장 분명한 것은 분포를 구한다는 점이고, 이와 관련해서는 분포개념, 즉 분포가 무엇인지, 왜 그것이 어려운지 살펴 보아야할 것이다. 둘째, 어떤 통계량의 분포를 구하는 것인데, 요약개념, 즉 통계량이란 무엇이고, 왜 그 문제에서는 그 통계량을 사용하는지, 그래서 통계적 추론 과정에서 그 정당화를 어떻게 이용, 판단할 것인지 개략적 계획이 세워져야 할 것이다. 셋째, 표본개념, 즉 그렇게 얻은 표집분포를 통해 통계적 추론을 했을 때, 그것이 과연 믿을만한 것인지, 확률적으로 얼마나 신뢰가능한지 이해가 되어야 한다는 것이다.

결국 이 세 가지 모두가 바르게 이해될 때 비로소 표집분포와 그 의미를 정확히 이해한 것이라고 한다면, 그리고 표집분포와 관련해서는 이 셋 중 어느 하나도 이해가 쉽지 않다면, 표집분포의 이해란 교사에게 있어서 가장 큰 교수학적 난제라고 할 것이다.

2. 통계적 핵심 개념과 표집분포

한국교육과정평가원의 연구보고서 '수학과 교육 내용 개선 방안 연구(윤현진 외, 2009)'에서는 일관성 있고 통합된 확률과 통계 교육과정을 구성하기 위한 개념틀로서 '도구적(분포) 개념', '타당성(요약) 개념', '신뢰성(표본) 개념'을 제안하고 있다. 이것은 학생의 인지발달 상황에 잘 대응하면서 통계적 개념이나 원리, 또는 논리를 포괄하는 통계적 방법론의 세 가지 개념이다.

다음에서는, 이 세 가지 개념에 대하여 알아보고 표집분포와의 관계를 알아보고자 한다.

가. 도구적(분포) 개념

일반적으로 통계적 방법론은 사회현상에서 발생되는 자료를 분석하기 위해 그 자료를 확률변수로 보고 그것의 분포를 사용해 총체적으로 파악하고자 한다. '도구적(분포) 개념'은 통계적 방법론이 현상을 분포에 의해, 즉 총체적으로 파악하는 목적으로 사용될 때, 이 방법을 이해하기 위해 필요한 개념을 의미한다.

통계학에서 분포개념은 총체로써 자료 집합을 볼 수 있게 하는 중요한 개념이다(Bakker, 2004). 여기서 분포는 여러 가지 현상에서 발생되는 자료의 속성이 다양할 때 그 속성 각각, 즉 범주의 빈도를 생각하는 개념으로써, 남주현(2007)은 현상을 전체적으로 파악하는 것을 목적으로 범주와 빈도를 연결 짓는 이러한 분포개념이 도구적 성격을 지닌다고 보았다. 즉 분포란 범주와 빈도의 순서쌍의 집합이라고 할 수 있다.

통계학에서는 여러 현상들이 '우연현상'을 지니고 있다고 가정한다. 우리가 인지할 수 있는 동일한 원인에 의해 다양한 결과를 얻을 수 있음을 인정할 때 우연현상이라고 판단하게 된다. 통계학은 우연현상을 연구 대상으로 삼아, 자료를 바탕으로 하는 귀납적 추론의 실제적 방법을 얻으며, 그 방법에 대한 과학적 정당화

에 관하여 연구하는 학문이다(이영하, 2006).

어떤 현상이 우연 속성을 지닌 것으로 판단될 때 그 현상을 다음과 같은 모형으로 설명할 수 있다(이영하, 2009).

$y = f(x) + \epsilon$

이것은 우연현상으로 보이는 y속에 잠재하는 필연성을 최대한 찾아내어 $f(x)$로서 설명하고 남은 부분인 ϵ만을 우연현상으로 설명한다.

우연현상에서 '변량의 변이성(variability)'은 우연현상을 설명하고자 얻은 법칙으로부터 어떤 예측을 얻으려 할 때 주된 장애로 작용하기 때문에, 이 변이성을 축소하여 더 정확한 예측을 도모하게 된다. 위의 모형에서 $f(x)$를 설명변인으로 볼 때 ϵ 은 우연변인으로 볼 수 있으며, 이 우연변인의 분포를 파악하여 비로소 현상을 분석할 수 있게 된다. 예컨데, 빈도개념, 분포의 시각적 표현, 상대도수의 비교, 표집분포 등은 모두 범주와 빈도의 의미를 연결 짓는 개념이며, 이러한 개념들을 바탕으로 도구적 개념이 형성되면 우연 변인의 분포를 이해하게 되는 것이다. 즉 우연 현상의 예측과 추론을 위해서는 분포라는 도구를 사용하는 것 때문에, 통계적 방법의 이해에서 분포개념이 반드시 필요한 이유이다.

나. 타당성(요약) 개념

'타당성(요약) 개념'은 추론하고자 하는 모집단의 정보에 타당한 표본의 정보가 무엇인가 또는 주어진 자료를 어떻게 요약해야 목적에 부합되는 요약이 될 것인가를 고려하고 판단할 때 필요한 개념이다.

타당성 개념의 기본수준개념으로 생활 속에서 일어나는 일의 상황이나 줄거리를 간단하게 정리하는 언어적인 요약과 사물이나 사건들의 분류활동 또는 범주화를 들 수 있다. 이는 수치적 정리를 이해할 수 있는 수준으로 발전해 나가는데, 목적에 맞게 수치적으로 정리를 한다는 것은 일반적으로 어떤 현상에서 수치적 정보에 초점을 맞추어 정리해 보는 것으로부터, 자료의 중심경향성이나 산포를 수치적으로 요약하는 것, 또는 자료대신 분포로써 주어진 경우에도 목적에 맞게 요약할 수 있는 것, 변동 가능한 자료값들에 대하여 설명변인과 우연변인에 근거한 모형을 만들 수 있는 것을 포함하게 된다(남주현, 2007).

타당성 개념은 주어진 자료, 또는 수집된 자료로부터 무엇을 계산할 것이며, 이 자료 값들이 서로 얼마나 비슷한가를 알아내기 위해 채택된 통계적 방법이 얼마나 타당한가를 판단하는 것과 관련된 문제로, Fisher(1925)는 자료를 축약할 때 무관한 정보를 제거하고 자료에 포함된 관련 정보 전체를 분리시키는 것이 채택된 통계적 과정의 목표라고 하였다. 따라서 통계적 방법론의 타당성개념의 핵심은 요약하려는 서술적 표현의 정량화와 관계된다.

모집단의 정보를 추출하고자 할 때는 어떻게 표본을 추출하여야 그로부터 요약된 통계량이 모집단을 잘 설명할 수 있을 것인가를 판정하게 된다. 표본과 관련된 통계량은 추정량으로서의 타당성, 즉 모수에 대한 합목적성이 중요한 문제가 되므로, 타당성 개념 관련 주제로 볼 수 있다. 그런데 표집분포는 바로 이러한 표본과 관련된 통계량이므로 표집분포는 타당성 개념을 이해하는데 매우 중요한 개념이라 할 수 있다.

앞의 모형 $y = f(x) + \epsilon$에서 $f(x)$는 통계분석의 (언어적으로 진술된)목적에 타당한 것이어야 하는데, 실제 정확한 $f(x)$ 식을 알기 어렵고, 또 x값의 영향의 정확성 역시 보장되지 않기 때문에, 실제 통계분석에서 얻어지는 $f(x)$의 추정 통계량 $\widehat{f(x)}$ 는 측정값 y들로부터의 계산결과에 의존하게 된다. 그런데, 다시 $\widehat{f(x)}$가 $f(x)$

와 관련하여 타당성이 있을 것이 요구된다. 본 연구는 앞의 표집분포의 분포개념 측면에 초점이 맞추어진 것이지만, "타당성 개념 없는 통계량"의 표집분포는 구해보았자 아무 가치가 없는 것이기 때문에, 타당성 개념은 표집분포 획득 후의 추론 과정에서 필수적인 개념이다.

다. 신뢰성(표본) 개념

'신뢰성(표본) 개념'은 변이성을 지닌 표본을 통해 믿을 수 있는 정보를 획득하는지를 비판적으로 이해하는 것이다. 따라서 표본 자체의 변이성을 분포개념을 통해 분석 관찰함으로써 이해하고, 논리적 측면에서 그와 같은 불확실성을 그대로 수용 담보하는 통계적 추론의 정당화 및 그 원리를 이해하는 것이 중요하다.

'표본의 변이성'은 표본이 모두 동일한 것은 아니며 모집단과 모두 닮은 것은 아니라는 것이다(Batanero et al., 1994). Moore(1992)에 따르면 통계적 사고의 핵심은 변이성이며 통계 교육의 궁극적 목적은 이 변이성과 자료를 현명하게 다루는 능력을 기르는 것이라고 하였다. 그러나 우정호(2000)는 통계적 사고의 불확정성, 부정확성, 무작위성이라는 인식론적 상태를 다루는 것은 결정론적 수학관과 연역적 방법으로 수학을 지도하고 학습해온 교사와 학생들에게 매우 어려운 일이기 때문에 통계적 사고의 핵심인 변이성의 이해와 교육이 쉽지 않으며, 따라서 이것이 현재 통계교육의 문제점이라고 지적하였다. 학생들은 형식적인 통계적 처리에만 익숙하고 이러한 통계적 사고의 특성에 대한 의식이 결여되어 있어 통계를 비판적 안목으로 바라보기 어렵게 되는 것이다.

'부분(표본)'으로 '전체(모집단)'를 예측하는 것은 오류가 발생할 가능성이 있다는 것은 직관적으로 알 수 있다. 하나의 통계가 표본으로부터 계산되었을 때(예를 들어, 표본평균), 이 수가 변이성을 지닌 것임을 인식하여 신뢰성에 의문을 가질 수 있다. 그러나 표집분포를 통해 그 계산 값의 신뢰성을 파악하는 것은 이해하기 쉬운 개념이 아니다.

표본개념에서는 반복되는 표본 추출에서 표본 자료 자체의 변화 가능성에 주목하는데, 이로부터 표본의 크기를 고려하는 '표집(sampling)'을 하게 되었다. 예컨대, 표본, 대표성, 표본의 크기, 표집개념 등은 통계적 추론의 신뢰성을 목적으로 사전에 고려해야 하는 개념들이다. 통계적 추론의 정당화를 위한 논리가 신뢰성 개념에 속한다는 것은 '가능성의 원리(likelihood principle)'로부터 알 수 있다.

'가능성의 원리'란 임의 표본에 의해 관측된 결과는 그와 같은 관측결과를 얻을 만한 충분한 확률적 이유가 있다(즉 확률이 높았을 것이다)고 보는 것을 말한다. 신뢰도는 불확실한 상황에서 가능성의 원리에 따라 판단 또는 의사결정을 하고 그와 같은 판단, 의사결정에 대한 확률적 자신감의 정도를 의미한다. 따라서 가능성 원리는 가설 검정 논리의 상식적 판단기준이 되며, 표본을 통한 정당화가 신뢰성 개념에 속함을 알 수 있다.

귀납추론에서는 일반화의 오류 가능성이 존재하는 것을 인지하는 것이 필요하다. 이경화(1996)에 따르면 이러한 인지는 귀납 논증에 의하여 일반화된 명제가 전적으로 참이라고는 할 수 없지만 확률적으로는 참이라고 할 수 있음을 이해한다는 것을 의미하는 것이므로 모평균의 구간추정에서 신뢰도와 신뢰구간을 이해하는 데 중요한 개념이 된다.

앞의 모형에 관한 논의에서 모평균 $m=f(x)$ 대신 표본평균 $\overline{X}=f(\widehat{x})$ 는 스스로 우연 변인을 지닌 것이기 때문에, 결국 이것의 확률분포, 즉 표집분포가 필요하고 이것이 확보된 후에는 타당성 개념과 가능성원리 등의 신뢰성개념을 토

대로 한 논리법칙에 의거하여 적절한 통계적 추론이 이루어져야 표집분포를 학습한 소기의 목적을 다 이루었다고 할 수 있을 것이다.

이상의 논의를 정리하면, 표집분포는 통계적 방법론의 세 가지 개념인 '도구적(분포) 개념', '타당성(요약) 개념', '신뢰성(표본) 개념' 등을 모두 포괄하고 있는 매우 중요한 개념이며, 통계 처리의 순서도 상에서 다음과 같이 표현할 수 있다.

모집단 ⇒ 표본자료(추출) ⇒ 통계량(분석/계산)

↑ (표본개념) ↑ (요약개념)

⇒ 표집분포(계산) ⇒ 추정/검정(추론)

↑ (분포개념) ↑ (표본개념)

본 연구는 위의 도표에서 "통계량"과 "표집분포" 사이의 표집분포의 "(분포개념)"의 이해를 위하여 시뮬레이션 프로그램을 개발하는 것이 주목적이며, 프로그램 제작 시 문제 상황설정 측면이나, 명령어 구성상의 배려를 통한 통계적 추론 능력 함양 수업 지원 차원에서 "(요약개념)" 이나 앞뒤의 두 "(표본개념)"을 고려하려는 것이다.

3. 우리나라 통계교육과정의 표집분포

우리나라 통계교육과정에서는 표본평균과 베르누이 독립시행에서의 표본합(성공 횟수)에 대한 표집분포를 다루고 있다. 전자는 중심극한정리의 형태로서 모평균에 관한 신뢰구간에서, 후자는 이항분포를 소개하는 과정에서 다루어진다. 하지만 표집분포에 대한 개념을 도입하지 않고, 단순히 두 개의 분포, 즉 이항분포와 정규분포에 대한 소개에 치중하게 됨으로써, 결과적으로 표집분포 개념 자체를 간단하게 설명하거나 아예 생략하고 마는 결과를 초래하고 있다.(윤현진 외, 2009).

만약 본 연구의 결과가 학교현장에서 사용된다면 그것만으로도 일단, 표집분포의 존재에 대한 인식 및 개념은 학교 통계 교실로 쉽게 확산될 수 있을 것이며, 전체 통계적 추론의 흐름 중에서 상실된, 적어도 한 개의, 중요한 교량이 회복됨을 의미한다.

그러나 요약개념의 타당성 부분은 권위적 정당화나 직관[7]에 호소하여 해결한다 해도, 표본개념의 신뢰성 관련 추론은 논리의 흐름이기 때문에 학생들이 반드시 스스로 이해하여야 할 부분이며, 컴퓨터의 역할 또한 미미할 것임으로, 이에 관한 교사의 역할이 특별히 강조될 필요가 있다. 물론, 이를 위해서는 교사 자신이 표집분포를 포함한 추론의 전 과정을 명확히 이해하고 있어야 할 것이다.

4. 컴퓨터 활용 수업 시 교사의 역할

컴퓨터와 정보통신을 수학교육에 활용할 때, 부정적 측면을 억제하고, 장점을 더 잘 드러내는 방안이 필요하며, 이를 위해 컴퓨터 활용에서 교사의 상황대처 능력은 매우 중요하다고 생각된다.

강문봉 외(2003)는 컴퓨터의 기술적 한계로 융통성 있는 자료 구현이 어려워 수준별 개별학습이 어렵고 획일적 방법을 강요할 위험이 있

7) "추론하려는 것이 모평균에 관한 것이니 표본평균이 당연히 타당한 통계량"이라고 하거나, "확률통계학자들에 의하면, 표본평균이 이 상황에서는 가장 좋다는 것이 증명되었다."라는 등의 초보수준의 언어 수리적 정당화를 뜻하며, 지나치게 상세한 타당성 논의를 시작하면 대학수준 통계학 추정론 수업으로 전환될 우려가 있음.

으며, 유연하거나 창의적이지 못하여 교사보다 다양한 방법을 제공할 수 없음을 지적하였다.

따라서 ICT를 활용하려는 교사는 어느 정도 프로그래밍 능력을 갖고 있어야 하며, 플래쉬나 비주얼베이직 등 개발 프로그램 중에서 저작도구를 표준으로 정해 ICT를 개발 보급하는 것이 바람직한데, 프로그램 개발자의 관점에서는 이것을 작성된 프로그램이 누구나 쉽게 알아볼 수 있는 소스 코드(source code)이어야 한다는 뜻이라고 이해할 필요가 있다.

특히 전문 프로그래머가 개발하였다면 그 뜻은 작성한 소스 코드(source code)를 공개하여 유지보수와 업그레이드가 가능하도록 해야 한다는 뜻으로 해석된다. 요즘은 소스 코드를 공개하고 공유하는 open source 운동이 활발하므로 교사들도 참여하여 공유하고 개발하는 것이 용이해졌다. 따라서 교사들이 교수·학습용 프로그램 개발에 관심을 가져야 할 것이다.

또 NCTM(2000)에서는 계산기나 컴퓨터가 기본적인 이해나 직관을 대체하는데 사용되어서는 안 된다면서, 오히려 그러한 이해와 직관을 촉진하는데 사용되어져야 하는데, 교사는 수업 중 이런 점에 경각심을 갖고 이해야 할 것이다.

컴퓨터에 대한 학생들의 맹신 역시 교사가 가장 주의 깊게 관찰해야할 사항이다. 컴퓨터가 실세계를 그대로 반영할 수 없다는 점, 가령 실세계에서는 연속적인 것이 컴퓨터에서는 이산적으로 표현될 수밖에 없음을 이해하는 것, 난수(random number)도 실제로는 의사 난수(pseudo-random number)[8]일 뿐이라는 점, 그래프나 도형의 나타내어지는 정도가 화면의 해상도에 따라 달라진다거나, 컴퓨터의 성능에 따른 여러 가지 제한이 불가피하다는 점 등을 학생이 충분히 이해할 수 있도록 지도할 필요가 있다.

교사는 컴퓨터를 활용할 때 드러날 수 있는 문제점을 과소평가해서는 안되며 이를 염두에 두고 수업에 활용해야 할 것이다. 컴퓨터는 만능해결사가 될 수 없음은 주지의 사실이다. 수학 학습에 수업이 집중될 수 있도록 교사의 적절하고 충분한 피드백이 요구되며 학생 개개인이 수업에 잘 따라가고 있는지 확인할 필요가 있다. 학습은 제대로 이루어 졌는지, 학습 목표를 달성하였는지 냉정하게 평가하는 것도 잊지 말아야 할 것이다.

5. 선행 연구 고찰

표집분포 개념을 고등학교 학생들이 이해하는 데는 많은 어려움이 따름은 주지의 사실이다. 특히 우리나라는 고등학교까지의 통계교육 과정에서 표집분포를 다루지 않기 때문에 학생들에게 표집분포의 개념을 이해시키기 위한 연구가 거의 이루어지지 않았다. 따라서 표집분포 개념 이해를 위한 시뮬레이션도 전혀 개발되어 있지 않다.

현재 우리나라 통계교육에서 사용되는 소프트웨어는 엑셀이 가장 많다. 엑셀 등을 활용한 통계 수업에 대한 많은 연구들이 이루어 졌는데(김동제, 박용범, 2001; 신보미, 이경화, 2006; 이진아, 2007; 장형도, 2006), 이러한 연구들도 마찬가지로 표집분포 개념 이해를 목적으로 하는 연구가 아닌 모의실험으로 확률을 계산해보거나 여러 가지 분포의 히스토그램과 그래프를 그려보는 수준에 그치고 있다.

김동제, 박용범(2001)의 엑셀을 활용한 통계

8) 컴퓨터에 의해서 만들어지는 난수. PC에서 난수를 만들 때에는, 컴퓨터의 현재 시각을 씨앗값(seed)으로 삼고 그 씨앗값을 수학적 알고리즘에 넣어서 난수를 생성한다. 따라서 컴퓨터 시계를 과거로 정확히 돌리면 똑같은 난수가 만들어지게 된다, 이는 결정적인 방식이므로 진정한 난수라고 할 수 없으며, 실제로는 아주 긴 주기를 가지고 있는 규칙적인 숫자의 열이다. 그러나 대개의 경우 주기가 매우 길기 때문에 난수라고 생각해도 무방하다. 컴퓨터에 의한 모의실험이나 통계 등에서 많이 사용된다.

수업의 흥미도 신장 방안에 대한 연구에 따르면 엑셀의 활용이 학생들의 수업에 대한 흥미를 불러일으키고 어느 정도 학습 효과를 가져오는 것으로 나타났으나, 학생들이 작성한 소감문을 보면 엑셀을 배우는 것이 힘들었으며 수학보다 엑셀 사용법이 더 어려워 도리어 수학 자체에 집중하지 못하는 모습을 보이는 경우도 있다. 이것은 엑셀이라는 수학 외적 환경(기술적 도구) 사용에 더 어려움을 느끼고 있는 것으로 매우 바람직하지 못한 것이다.

학생들이 엑셀 고유의 용어와 사용법에 익숙해지기까지는 많은 시간이 소요된다. 또한 엑셀의 기능에 따라 수업 화면이 계속 바뀌게 되어 수업 내용을 한 화면에 일관되게 진행할 수 없는 것도 학생들이 수학적 개념의 발전 과정을 학습하는데 방해가 된다.

통계용 소프트웨어인 미니탭을 활용한 수업에 대한 연구(노지희, 2009)도 있으나, 히스토그램을 그려보는 수준이며, 무엇보다도 미니탭이나 엑셀이라는 프로그램에 의존하기 때문에 해당 소프트웨어를 반드시 구매해야하는 문제점이 있다.

우리나라와 달리 미국에서는 표집분포 개념의 중요성에 대하여 인식하고 있으며, 여러 연구자들이 학생들의 표집분포 개념 이해를 돕기 위해 시뮬레이션 개발을 제안하였다(Behrens, 1997; Davenport, 1992;, Glencross, 1998;, Rubin & Bruce, 1991; Schwarz & Sutherland, 1997; Simon, 1994). 그러나 개발된 시뮬레이션이 제대로 활용되고 있거나 활용할 수 있도록 제공되어 있는 것은 거의 없는 실정이다(delMas, 2004). 이에 delMas는 2001년 대학생들을 대상으로 한 통계 수업을 위해 표집분포 학습을 위한 'Sampling SIM'을 개발하였다. 그러나 이것을 우리나라 고교수업에 직접 적용하기에는 교육과정 자체가 다르다는 점, 사용되는 용어가 고교수준을 넘어선다는 점, 언어가 다르기 때문에 결과해석이 용이하지 않은 점, 화면 구성이 조작적이지만 수업을 이끌어 나가기에 복잡하다는 점 때문에 적절하지 않다.

따라서 고등학교 수준의 학생들이 쉽고 편하게 사용할 수 있으며, 수업 내용을 따라가기에 무리가 없고, 값비싼 소프트웨어에 의존하지 않는 '표집분포 시뮬레이션'을 개발하여 학생들이 표집분포의 개념을 이해하도록 돕는 것이 필요하다.

III. 표집분포 시뮬레이션

1. 표집분포 시뮬레이션

가. 개발 방법

정보과학에서 시뮬레이션이란 실제 상황을 컴퓨터 내에서 인위적으로 구현해보는 방식을 말한다. 가령 어느 모집단에서 크기 100인 표본을 구하여 평균을 구할 때, 가능한 평균값의 분포를 얻기 원한다면, 이 문제의 답을 얻기 위한 시뮬레이션 방법은, 컴퓨터 내에 똑같은 모집단을 만들어 놓고, 인공적인 모집단에서 크기 100인 표본을 반복해서(무한히 반복할 수 없으므로 매우 큰 수의 유한 번 반복함) 얻어 매번 평균값을 구하고, 그 평균값들을 가지고 도수분포표를 얻는 방식으로, 표본평균의 분포를 얻는 것이다.

표집분포 시뮬레이션은 바로 이런 형태의 프로그램을 구성하려는 것이다.

나. 개발의 특징

국내의 연구에서는 표집분포 지도에 관한 것이 없으나 미국에서 대학생을 대상으로 한 교육용 컴퓨터 소프트웨어가 존재한다고 하였다.

그러나 이런 소프트웨어가 보급이 미미하고 사용이 활성화되지 않은 가장 중요한 이유는 얻어진 표집분포를 실제 어떻게 활용하는 지, 즉 그 활용의 과정인 통계적 추론이 용이하지 않기 때문이라고 추측하였다. 왜냐하면 앞서 분포개념만이 아니라 요약개념과 표본개념 역시 표집분포를 이용한 통계적 추론에서는 중요한데, 이들 역시 이해가 쉽지 않다고 본 연구자는 판단하였기 때문이다.

따라서 본 연구를 통해 개발하려는 소프트웨어는 단순히 표집분포만을 지도하려는 프로그램이라기보다는 표집분포를 이해하여 바른 통계적 추론을 할 수 있도록, 실제 표집분포의 역할과 추론적 기능을 체험할 수 있도록 하려는 것이다. 즉 통계적 추론이 매우 강조된 표집분포 지도 프로그램이라고 할 수 있다.

다. 개발 방향

'표집분포 시뮬레이션'은 앞서 언급한 그런 관점에서 학생들의 접근이 용이하도록 개발되었다. 세부적인 '표집분포 시뮬레이션'의 개발 방향은 다음과 같다.

1) 다른 소프트웨어나 프로그램을 필요로 하지 않게 한다.

기존의 통계교육에서 컴퓨터의 활용은 주로 엑셀을 사용하거나 통계용 소프트웨어(SAS, SPSS, 미니탭) 등 값비싼 소프트웨어를 이용한 것들이 대부분이다. 이런 다른 프로그램에 의존하는 소프트웨어는 경제적으로나, 사용의 간편성, 작동의 즉시성 측면에서 매우 제한적일 수밖에 없다. 또 사용자 입장에서도, 해당 소프트웨어 사용법을 알아야 한다는 불편함이 따른다.(김동제, 박용범(2001), 신보미, 이경화(2006), 이진아(2007)) 본 연구의 결과는 고등학교 통계 수업 어느 곳에서나 간단한 준비로 즉시 사용할 수 있어야 한다.

2) 한글을 사용하여 이해하기 쉽고, 사용법이 편리하다.

교육용 컴퓨터 프로그램이 너무 복잡하면 도리어 산만해져서 학생들이 수업에 집중하는 것을 방해한다. '표집분포 시뮬레이션'의 사용자 환경(user interface)은 한글과 고교수준의 용어를 사용하되, 프로그래밍 언어는 Visual Basic을 사용한다.

3) 표집분포의 의미를 통계적 추론과 관련지어 이해하도록 한다.

표집분포 뿐 아니라 표집분포와 관련된 신뢰구간, 중심극한정리, 이항분포의 정규근사 등도 함께 다루어 표집분포를 통한 통계적 추론의 경험을 통해 추론 능력을 기르고, 표집분포의 이해가 왜 중요한지 깨달을 수 있도록 한다.

4) 복잡한 계산은 피하되 필요한 경우 직접 계산하도록 구성하였다.

표본평균의 표집분포를 이해하려면, 그것이 표본평균들의 분포임을 알아야 한다. 그런데 평균 구하기를 계산상의 복잡성을 이유로 컴퓨터에게 모두 맡겨버리면, 무엇의 분포를 구한 것인지 결국 설명만으로 듣고 암기하는 방식이 된다. 이것을 막기 위해 적어도 몇 번은 직접 평균을 구해보도록 하여 자신이 얻은 분포가 평균들의 분포임을 느끼도록 할 필요가 있다. 이렇게 컴퓨터와 서로 질의 응답하는 과정을 거치게 하는 명령어 구성이 필요하다. 또한 제일 하단의 '상태표시영역'에 현재 시뮬레이션의 상태를 표시하여 학생들이 시뮬레이션의 진행 과정을 보고 알 수 있도록 하여, 직접 활동하지는 못하더라도 무엇을 공부하고 있는 지 잊지 않도록 유의한다.

5) 학생들이 수업 진행 과정을 따라가기 쉽다.

개념의 발전 과정이 자연스럽게 이루어지도록, 각 차시의 수업이 한 화면에서 순차적으로 진행되도록 하고, 학생들이 수업 과정을 한눈에 직접 확인할 수 있게 한다.

6) 교사와 학생간의 유기적인 상호작용이 가능하도록 하였다.

통계적 추론 과정에서 타당성 개념이나 신뢰성 개념에 의한 통계적 추론 내용은 컴퓨터만으로는 지도가 불가능하다고 생각한다. 따라서 각 차시마다 학생들이 컴퓨터만 가지고 혼자 학습하는 상황이 아니라, 교사의 적극적인 배경 설명과 개입이 있을 것이라는 가정 아래서 프로그램을 구성한다.

7) 프로그램 사이즈가 작고 실행시키는데 많은 시간이 소요되지 않는다.

이 프로그램은 공용 소프트웨어임으로 상호복제 전파가 용이하여 수업준비가 간단하고, 상대적으로 열악한 환경에서도 쉽게 작동되어야 한다. 또, Visual Basic 프로그램 구사가 가능한 교사는 간단히 프로그램을 고칠 수 있어야 한다. 따라서 '표집분포 시뮬레이션'은 실행파일의 크기가 작고, 실행 시 컴퓨터의 메모리를 많이 차지하지 않아야 한다. 또 실행하는데도 많은 시간이 소요되지 않도록 한다.

2. 세부적 구현 방법

가. 정규모집단

1) 정규모집단 생성 (1차시, 2차시)

구간 (0,1)에서 두 개의 난수 u_1과 u_2를 발생시킨 후, 'Box & Muller방법'[9]에 대입하여 표준정규분포를 따르는 두 수 z_1과 z_2를 구하고, $x=\mu+\sigma z$에 대입하여 정규분포 $N(\mu, \sigma^2)$을 따르는 두 수 x_1과 x_2를 생성한다.

위의 방법으로 정규분포를 따르는 수가 2개 생성되므로 위의 과정을 10,000회 반복하여 크기 20,000인 모집단을 생성한다. 생성된 모집단은 컴퓨터의 메모리에 배열 형식으로 저장해둔다.

2) 표본추출하기 (1차시, 2차시)

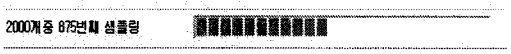
위의 과정에서 만들어진 모집단으로부터 주어진 크기의 표본을 추출한다. 표본을 추출할 때, 배열에서 바로 추출하게 되면 중복 추출 여부를 확인하는 작업을 수행하는 번거로움이 있으므로 Visual Basic의 컬렉션(Collection)을 사용하여 중복 추출이 되지 않도록 하였다. 컬렉션은 배열과 달리 컬렉션에 저장되어 있던 항목(Item)을 제거(Remove)하면, 컬렉션의 남은 항목들의 인덱스(Index)가 재조정되어 비복원 추출이 용이해진다.

표본추출 과정은 다음과 같다. 먼저 메모리에 배열로 저장되어 있던 모집단을 컬렉션에 다시 저장(Add)한 후, 컬렉션에서 표본 크기만큼 랜덤하게 항목을 비복원으로 추출한다. 다시 컬렉션을 초기화하고 모집단을 컬렉션에 저장한 후, 표본 크기만큼 랜덤하게 항목을 비복원으로 추출한다. 위의 과정을 표본추출 횟수만큼 반복한다. 그러므로 한 번 표본을 추출할 때에는 비복원추출을 하고, 다음 표본을 추출할 때에는 모두 복원시켜 다시 표본을 추출하는 복원추출을 하게 된다.

표본을 자동으로 추출하게 되면 많은 수의 표본을 추출하고 적절한 연산을 해야 하므로 실행시간이 어느 정도 소요된다. 이 경우 하단의 '상태표시영역'에 작업진행표시 막대를 사용하여 어느 정도 실행되고 있는지 상태를 보여주도록 하였다. 아래 [그림 III-1]은 2,000회의

9) $z_1 = \sqrt{-2\times\log(u_1)} \times Cos(2\times 3.1415926\times u_2)$
$z_2 = \sqrt{-2\times\log(u_1)} \times Sin(2\times 3.1415926\times u_2)$

표본추출횟수 중 현재 895번째 표본들을 추출 중임을 나타낸다.

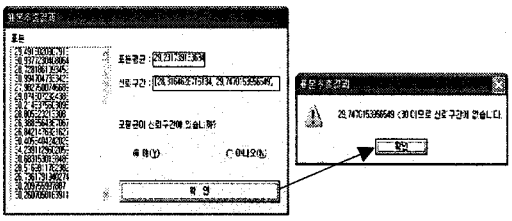


[그림 III-1] - 표본추출(자동) 시 작업진행표시 막대

3) 계획신뢰도와 실제신뢰도 (1차시)

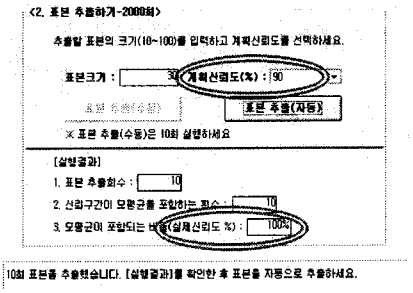
신뢰도란 표본을 추출하여 표본의 평균을 구하고 그로부터 계산된 신뢰구간이 모평균을 포함할 것이라고 기대하는 정도를 의미한다. 1차시에서는 신뢰도가 두 번 나타나는데, 첫 번째는, 위와 같이 신뢰구간을 구하기 위한 이론적 신뢰도이다. 두 번째는 구해진 신뢰구간이 모평균을 포함하는 횟수를 표본추출횟수에 비례하여 확인해본 결과적 신뢰도이다. 이 두 가지를 구별하기 위하여, 전자는 '계획신뢰도', 후자는 '실제신뢰도'로 표현하였다.

수동으로 표본을 10회 추출할 경우, 계획신뢰도가 90%일 때, 매 수동추출 시마다 다음 [그림 III-2]와 같은 화면이 pop-up 된다. 학생들이 옳은 값을 선택하지 못하였을 경우엔 다시 pop-up 화면으로 설명하여 주고 '표본추출(수동)'으로 돌아간다. 추출된 표본들은 ListBox 안에 출력되도록 하였다.



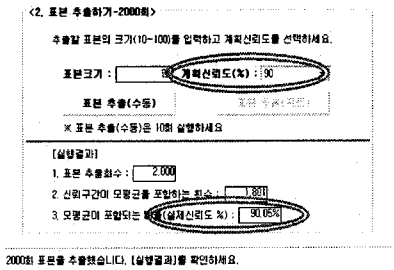
[그림 III-2] 1차시 표본추출(수동) 시 결과 pop-up창

아래 [그림 III-2]은 크기 30인 표본을 수동으로 10회 추출 완료했을 때, 계획신뢰도에 따라 계산된 각각의 신뢰구간이 모평균을 몇 번이나 포함했는가를 알려주고, 그에 따른 실제신뢰도를 알려주는 것이다. 계획신뢰도는 90%이지만, 실제 표본을 추출했을 경우엔 10회 모두 모평균을 포함해 실제신뢰도가 100%가 되었음을 알 수 있다.



[그림 III-3] 1차시 표본추출(수동) 결과

아래 [그림 III-4]는 표본을 자동으로 2,000회 추출했을 경우의 결과화면이다. 실제신뢰도가 90.05% 로 수동추출보다 계획신뢰도(90%)에 매우 근사함을 알 수 있다.



[그림 III-4] - 1차시 표본추출(자동) 결과

4) 수동으로 표본추출하기 (2차시)

2차시는 표본의 크기마다 수동으로 10회 표본을 추출하고 각 추출 시마다 표본평균을 계산하여 보여줌으로써 표본평균의 표집분포를 이해하도록 하였다. 1차시와 같이 추출된 표본들은 ListBox 안에 출력되도록 하였다.

5) 히스토그램 그리기 (2차시)

히스토그램의 계급의 수는 일반적으로 H. A. Sturges 가 제안한 아래 방법을 사용하여 구한다.

계급의 수 $= 1 + [\log_2 n] \fallingdotseq 1 + 3.3\log n$

그러나 H. A. Sturges의 방법에 따라 계산해 보면 모집단의 크기가 20,000인 경우에도 계급의 수가 15 밖에 되지 않기 때문에 히스토그램을 그렸을 때 명시성이 떨어진다. 특히, 표집분포 히스토그램의 경우 표본평균들이 모평균 근처에 집중되므로 학생들이 학습하기에 좋은 그래프가 그려지지 않는다. 따라서 구간을 더 세분하여 50개로 나누어 학생들이 시각적으로 구분하기에 용이하도록 하였다.

가) 정규모집단의 히스토그램

정규분포 $N(\mu, \sigma^2)$을 따르는 정규모집단의 히스토그램은 구간 $[\mu - 4\sigma, \mu + 4\sigma]$ 안에서 그리도록 하였다. 모집단 내의 최소값과 최대값을 각각 양끝 값으로 설정하여 그릴 경우, 우연히 아주 작은 값이나 매우 큰 값이 있게 되면 그래프의 모양이 한쪽으로 치우치게 된다. 정규분포에서 μ와 σ에 따른 확률은 다음과 같으므로 (경험규칙 (Empirical Rule), Mendenhall at al., 1990)

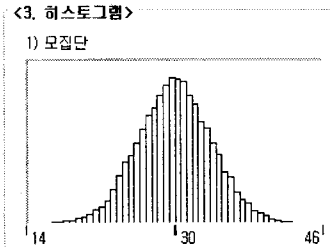
① $P(\mu - \sigma \le X \le \mu + \sigma) = 0.683$

② $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.954$

③ $P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997$

구간 $[\mu - 4\sigma, \mu + 4\sigma]$은 모집단 대부분을 포함한다고 보아도 무방하며, 따라서 학생들이 학습하는 데는 별 무리가 없다.

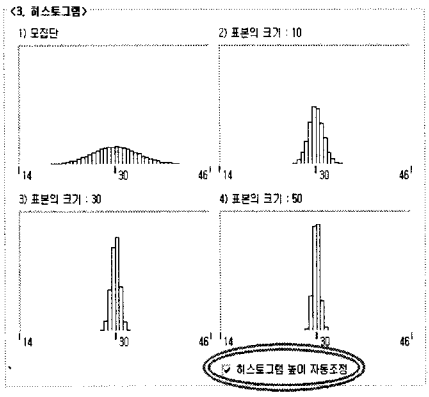
모평균 30, 모표준편차 4인 정규모집단에서 $\mu - 4\sigma = 14$, $\mu + 4\sigma = 46$이므로 모집단의 히스토그램은 다음 [그림 III-5]와 같이 그려진다. 가로축의 좌표값은 왼쪽 끝에 $\mu - 4\sigma$값 (=14), 중앙에 모집단의 평균값 (=30), 오른쪽 끝에 $\mu + 4\sigma$값 (=46) 으로 설정한다.



[그림 III-5] 평균 30,표준편차 4인 정규모집단의 히스토그램

나) 표집분포의 히스토그램

표집분포의 히스토그램에서 가로축의 좌표값은 모집단 히스토그램의 좌표값과 동일하게 하여 두 그래프의 비교가 용이하도록 하였다. 다음 [그림 III-6]은 2차시에서 모평균 30, 모표준편차 4인 정규모집단 히스토그램과 표본의 크기가 각각 10, 30, 50 일 때의 표집분포 히스토그램이다.



[그림 III-6] 정규분포 $N(30, 4^2)$를 따르는 모집단과 표집분포 히스토그램

위의 히스토그램에서 표집분포의 히스토그램들은 모집단에 비해 표본의 크기가 커질수록 평균을 중심으로 모이면서 높아져 뾰족해짐을 알 수 있다. 표집분포는 모집단으로부터 추출한 표본들의 평균들의 분포이므로 평균을 중심으

로 밀집되고, 따라서 표집분포의 평균은 모집단과 같으나 분산은 n에 반비례($\frac{\sigma^2}{n}$) 함을 알 수 있다. 그러나 학생들이 이해하기는 어려우므로 위와 같이 시각적으로 확인하도록 함으로써 표집분포의 개념을 쉽게 이해하도록 하였다.

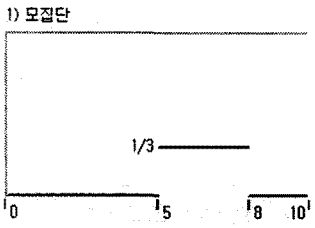
다) 히스토그램 높이 자동조정

히스토그램 하단에 있는 '히스토그램 높이 자동조정' 체크박스를 선택하고 히스토그램을 그리면 이미 그려져 있던 히스토그램과 새로 그려질 히스토그램의 y축 값을 비교한 후 적당한 비율로 축소 또는 확대하여 높이를 조정하도록 하였다(그림 6 참고). 높이를 자동으로 조정하려면 히스토그램을 그리기 전에 미리 체크박스를 선택하여야 한다.

나. 일양모집단 (3차시 일양분포)

1) 일양모집단의 히스토그램

일양분포 $U(a, b)$를 따르는 모집단의 히스토그램은 일반적인 일양분포의 확률밀도함수를 그리고 함수값($=\frac{1}{b-a}$)을 그래프 좌측에 표시하여 일양분포의 확률밀도함수의 의미를 잘 이해하도록 하였다. [그림 III-7]은 일양분포 $U(5, 8)$을 따르는 모집단의 확률밀도함수 그래프이다.



[그림 III-7] 일양분포 $U(5, 8)$을 따르는 모집단의 확률밀도함수 그래프

2) 수동으로 표본추출하기

3차시에서 일양모집단은 1차시와 2차시 정규모집단과 달리 따로 생성하지 않고 바로 표본을 추출하도록 한다. 일양분포 $U(a, b)$를 따르는 일양모집단에서의 표본추출 방법은 다음과 같다.

구간 (0,1)에서 난수 u를 발생시킨 후, 공식 $x=u\times(b-a)+a$에 대입하여 일양분포 $U(a, b)$를 따르는 x를 생성한다. 표본의 크기에 따라 난수를 생성한 후 이들의 평균을 구한다. 이 과정을 표본추출횟수 만큼 반복한다.

표본의 크기가 10일 경우엔, 표본평균들 표집분포임을 개념적으로 알 수 있도록 추출된 표본들의 평균을 구해서 확인하는 과정을 거치도록 하였다. 소수점이하 자리수가 너무 많으면 계산이 용이하지 않으므로 계산을 위한 표본들의 값은 소수점이하 셋째자리까지 반올림하여 나타내도록 하였고, 그 합도 마찬가지로 소수점이하 셋째자리까지 구하여 학생들이 직접 입력하도록 하였다. 만약 틀린 값을 입력하였을 경우엔 정답을 pop-up창으로 띄워주고 다시 표본을 추출하는 과정으로 돌아가도록 하여 교사가 피드백을 해주느라 시간을 소비하지 않도록 하였다.

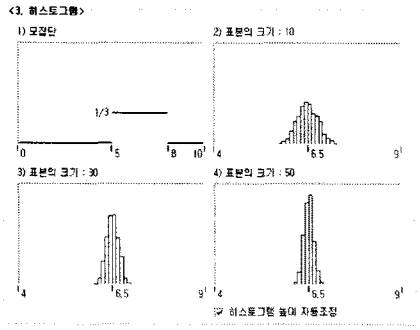
3) 자동으로 표본추출하기

자동으로 표본을 추출하는 경우엔 위의 수동으로 추출하는 과정을 표본추출횟수만큼 반복한 후, 얻어진 평균들로 히스토그램을 그린다. 학생들이 직접 평균을 구해 입력하는 과정은 포함하지 않는다.

4) 표집분포 히스토그램

표본평균들이 구해지면 정규분포 시 표집분포의 히스토그램과 같은 방법으로 히스토그램을 그린다. 단, 일양분포의 표본평균들의 표집분포이므로 가로축의 구간은 $[a-1, b+1]$ 로 설정한다. 계급의 구간은 정규모집단의 표집분포 히스토그램의 경우와 마찬가지로 50개로 나누어 히스토그램을 그리도록 하였다. 가로축의 좌표값은 왼쪽 끝에 $(a-1)$ 값, 중앙에 $(\frac{a+b}{2})$

값 , 오른쪽 끝에 $(b+1)$ 값으로 설정한다. 아래 [그림 III-8]은 일양분포 $U(5, 8)$을 따르는 모집단과 표본 크기가 각각 10, 30, 50 인 표집분포의 히스토그램이다. 일양분포의 표집분포 히스토그램은 정규분포와 마찬가지로 중심으로 모이면서 높아져 뾰족해짐을 알 수 있다.

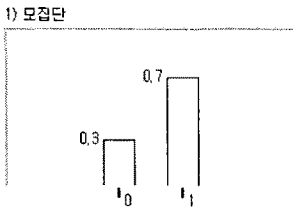


[그림 III-8] 일양분포 모집단과 표집분포 히스토그램

다. 이항모집단

1) 이항모집단의 히스토그램 (3차시)

이항분포 $B(n, p)$는 어떤 사건이 일어날 확률이 p인 베르누이 시행을 n회 반복한 결과의 합이므로 모집단의 분포는 0과 1에 대한 막대그래프 모양이 될 것이다. 확률 p가 0.7인 경우 이항모집단의 막대그래프는 다음 [그림 III-9]와 같다.



[그림 III-9] 확률 0.7 인 경우 이항모집단의 막대그래프

2) 수동으로 표본추출하기

이항분포 $B(n, p)$는 어떤 사건이 일어날 확률이 p인 베르누이 시행을 n회 반복한 것이므로, 이항분포 $B(n, p)$을 따르는 모집단의 표본은 베르누이 시행 결과의 합으로 구할 수 있다.

가) 선택된 p 값과 구간 (0,1)에서 추출된 난수 u를 사용하여 표본을 n개$(x_1, x_2, x_3, \cdots, x_n)$ 추출한다. 예를 들어, 베르누이 분포는 0 과 1 만 갖는 분포이므로, $p=0.7$ 인 경우, 추출된 u의 값에 따라, $\begin{cases} u \le 0.7 & \Rightarrow \ x_i = 1 \\ u > 0.7 & \Rightarrow \ x_i = 0 \end{cases}$ 의 값을 갖도록 한다(단, $i=1, 2, 3, \cdots, n$).

나) 추출된 표본들의 표본합$(=\sum x_i)$을 구한다. 따라서 표본합의 값의 범위는 0 ~ n이 된다. 4차시에서는 이 표본합의 분포를 생각한다.

다) 위의 표본합을 n으로 나누어 표본평균 $(=\frac{\sum x_i}{n})$을 구한다. 따라서 표본평균 값의 범위는 0 ~ 1이 된다. 3차시에서는 이 표본평균의 표집분포를 학습하게 된다.

라) 위의 과정 i) ~ iii)을 2,000회(표본추출횟수) 반복하여 표본평균 2,000개를 구한다.

표본의 크기가 10일 경우엔 일양분포와 마찬가지로 추출된 표본의 평균을 학생들이 직접 입력해보도록 하고 입력한 평균값이 틀린 경우 정답을 pop-up 화면으로 보여준다.

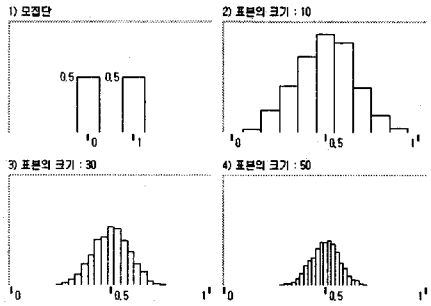
3) 자동으로 표본추출하기 (3차시 이항분포)

자동으로 표본을 추출하는 경우엔 위의 수동으로 추출하는 과정을 표본추출횟수만큼 반복한 후, 얻어진 평균들로 히스토그램을 그린다. 학생들이 직접 평균을 구해 입력하는 과정은 포함하지 않는다.

4) 표집분포 히스토그램 (3차시 이항분포)

표본평균들이 구해지면 정규분포에서 표집분포의 히스토그램과 같은 방법으로 히스토그램을

그린다. 표집분포의 히스토그램은 이항분포의 그래프의 연속성 보정을 사용하여 그리도록 하였다. 가로축의 계급의 수는 표본의 크기와 같게 되나, 표본평균의 표집분포이므로 가로축의 값은 $0 \sim 1\,(\frac{0}{n} \sim \frac{n}{n})$이 된다. [그림 III-10]은 $p = 0.5$일 때, 표본 크기가 각각 10, 30, 50 인 표집분포의 히스토그램이다.

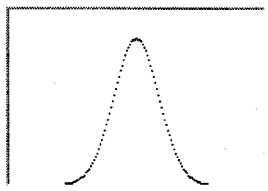


[그림 III-10] 이항분포 모집단과 표집분포 히스토그램

5) 이항분포의 정규근사 (4차시)

가) 정규근사 그래프

이항분포 $B(n, p)$의 정규근사 $N(np, npq)$의 그래프는 가로축 구간$[np-4\sqrt{npq},\ np+4\sqrt{npq}\,]$를 10,000개로 나누어 각 구간의 x값을 정규분포의 확률밀도함수에 대입하여 얻은 값으로 히스토그램을 그린 후 히스토그램 윗부분의 선분만 나타나도록 하여 마치 점선으로 연결된 정규분포 그래프처럼 보이도록 하였다[그림 III-11]



[그림 III-11] 이항분포의 정규근사 그래프

나) 정규근사와 표본합의 비교

정규근사의 경우 평균, 분산과, 표본합들의 평균, 분산을 계산하여 얼마나 근사한지 확인할 수 있도록 하였다. 베르누이 분포에서 사건이 일어날 확률 $p = 0.5$일 때, 정규근사와 이항분포의 표본합의 평균과 분산은 [그림 III-12]와 같다.

<2. 정규근사와 표본합의 비교>

표본크기	평 균		분 산	
	정규근사(np)	표본합	정규근사(npq)	표본합
5	2.500	2.495	1.250	1.224
15	7.500	7.497	3.750	3.811
30	15.000	15.033	7.500	7.548

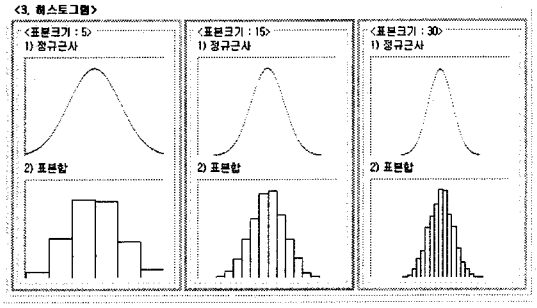
<n : 표본크기, p : 확률, q=1-p>

[그림 III-12] 정규근사와 표본합의 평균과 분산 비교

다) 표본합의 히스토그램

표본합의 그래프는 이항분포 그래프의 연속성 보정과 비슷한 방법으로 그리도록 하였다.

아래의 [그림 III-13]은 베르누이 분포에서 확률 p가 0.5 인 경우 정규근사와 표본합의 히스토그램을 비교한 것이다.



[그림 III-13] 확률이 0.5일 때, 정규근사와 표본합의 히스토그램

라. 기타 세부 고려 사항

1) 숫자 표현 규칙

표집분포 시뮬레이션에서 보여 지는 숫자는 3차시에 학생들이 직접 평균을 계산하는 부분을 제외하고 컴퓨터가 표현할 수 있는 범위 그대로 사용하였다. 학생들이 직접 계산하는 부분에서는 소수점이하 셋째자리까지 반올림한 값으로 계산하도록 하였는데, 계산 결과보다 그

의미를 이해하는 것이 목적이므로 복잡한 계산은 피하도록 하였다.

정규모집단 생성 시 평균값은 입력 가능한 범위를 8자리로 한정하였다. 이것은 모집단 히스토그램을 그릴 때 가로축 좌표 표현에 문제가 되지 않도록 하고, 너무 큰 값을 입력하여 학습에 방해가 되지 않도록 하기 위함이다.

2) 상태표시영역

각 차시 화면의 제일 하단은 '상태표시영역'으로, 현재 시뮬레이션이 수행하고 있는 과정이 무엇이며 해당 과정의 완료 여부와 진행 상태를 알 수 있도록 하였다.

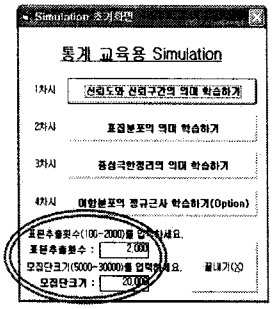
3) '새로시작', '끝내기' 버튼

각 차시 상단의 '새로시작' 버튼을 클릭하면, 그동안 진행되었던 모든 과정을 처음부터 다시 실행할 수 있다. 각 차시 상단의 '끝내기' 버튼을 클릭하면 해당 차시를 끝내면서 수업 중이던 창이 사라진다. 전체 시뮬레이션을 마치려면 '통계교육 시뮬레이션 초기화면'의 끝내기 버튼을 클릭하도록 한다.

3. 사용자 환경 및 수업 과정

가. 시뮬레이션 초기화면

표집분포 시뮬레이션을 실행시키면 다음과 같은 초기화면이 구동된다[그림 III-14]



[그림 III-14] 시뮬레이션 초기 화면

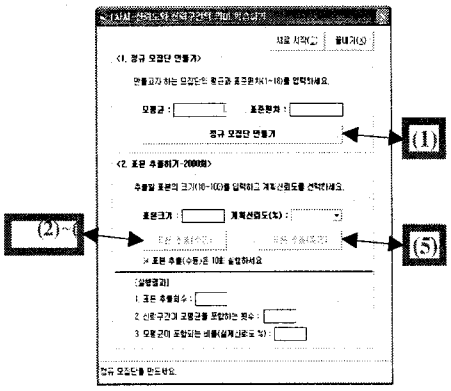
각 차시의 제목을 클릭하면 해당 차시의 학습내용이 구동된다. '표본추출횟수'와 '모집단크기'를 학습 시작하기 전에 변경 가능하도록 하였다. 각 차시별로 처음 시작할 때 변경 가능하고, 해당 차시를 실행하는 동안에는 변경할 수 없다. 너무 작은 값으로 설정하면 원하는 결과를 얻지 못할 수도 있으므로 가능하면 변경하지 않고 사용하도록 한다.

나. 신뢰도와 신뢰구간의 의미 학습하기

1) 학습목표

모평균과 표준편차를 입력하여 모집단을 만들고, 그로부터 표본크기 만큼 추출한 표본들의 표본평균과 계획신뢰도로부터 얻어진 신뢰구간에 모평균이 포함되는지 여부를 확인해봄으로써 신뢰도와 신뢰구간의 의미를 이해할 수 있다.

2) 1차시 초기 화면 [그림 III-15]



[그림 III-15] 1차시 초기 화면

3) 수업 과정 ([그림 III-15] 참고)

가) '모평균'과 '모표준편차'를 입력하고 '정규 모집단 만들기' 버튼을 클릭한다. 모표준편차의 범위는 1~18이므로 잘못 입력 시 에러메시지가 pop-up 된다. 하단의 상태표시영역에 '정규 모집단을 만들었습니다.'라는 메시지가 뜨면 다음 과정을 진행한다.

나) 추출할 표본의 크기와 신뢰도를 입력하

고 '표본추출(수동)'을 클릭한다.

다) 추출된 표본의 평균과 신뢰구간을 확인하고 모평균이 신뢰구간에 포함되는지 여부를 확인한다. 잘못 입력 시 에러메시지가 pop-up된다.

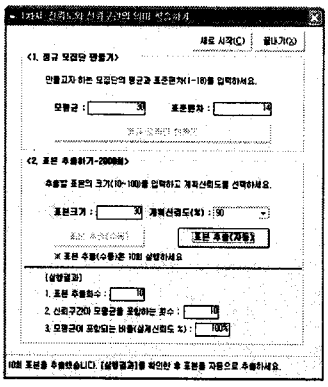
라) 위의 '표본추출(수동)' 과정을 10회 반복하고, 그 결과를 아래의 [실행결과] 에서 확인한다. 교사는 실행 결과를 바탕으로 하여 '계획신뢰도'와 '실제신뢰도'의 차이를 설명한다.

마) '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하단의 상태표시영역에 작업진행 표시 막대가 사라지고 '2,000회 표본을 추출했습니다.'라는 메시지가 뜨면, [실행결과]를 확인한다.

바) 교사는 학생들이 신뢰도의 의미뿐만 아니라 '계획신뢰도'와 '실제신뢰도'가 다른 이유를 정확히 이해하고 있는지 발문 등으로 반드시 확인한다.

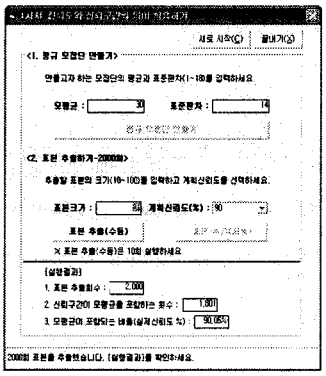
사) 표본추출을 다시 하려면 '표본추출(수동)'이나 '표본추출(자동)'을 클릭하고, '정규 모집단 만들기'부터 다시 하려면 상단의 '새로시작' 버튼을 클릭한다.

4) 표본추출(수동) 10회 실행 결과 화면



[그림 III-16] 1차시 표본추출(수동) 실행 결과 화면

5) 표본추출(자동) 실행 결과 화면 [그림 III-16]



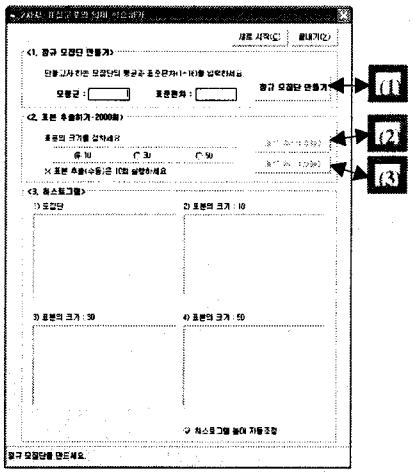
[그림 III-17] 1차시 표본추출(자동) 실행 결과 화면

다. 표집분포의 의미 학습하기

1) 학습목표

모집단을 만들고 표본을 추출해봄으로써 표집분포가 표본평균들의 분포임을 알 수 있으며, 모집단과 추출하는 표본의 크기에 따른 표집분포의 히스토그램을 비교함으로써 표집분포의 평균과 분산, 나아가 표집분포의 의미를 이해할 수 있다.

2) 2차시 초기 화면 [그림 III-18]



[그림 III-18] 2차시 초기 화면

3) 수업 과정 ([그림 III-18] 참고)

가) '모평균'과 '모표준편차'를 입력하고 '정규 모집단 만들기' 버튼을 클릭한다. 모집단의 히스토그램을 확인한다.

나) 추출할 표본의 크기를 선택하고 '표본추출(수동)'을 클릭하여 추출된 표본들과 표본평균을 확인한다. 이 과정을 10회 반복한다.

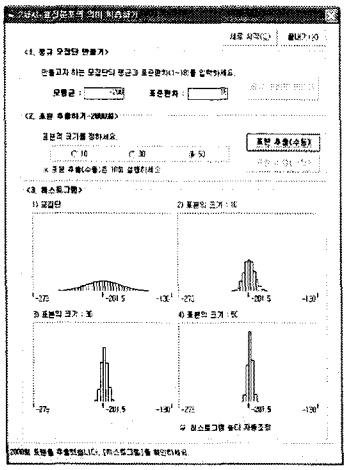
다) '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하단의 상태표시영역에 작업진행표시 막대가 사라지면 표집분포의 히스토그램을 확인하고, 모집단의 히스토그램과 비교해본다.

라) 표본의 크기를 변경하여 위의 (2)~(3) 과정을 진행한 후 모집단과 표집분포의 히스토그램을 비교하여 본다.

마) 교사는 표집분포의 히스토그램이 중앙으로 모이면서 더 뾰족하게 솟았음을 강조하면서 표집분포의 평균은 모집단과 비슷하나 분산은 표본의 크기(n)에 반비례함을 학생이 이해할 수 있도록 한다.

바) 표본추출을 다시 하려면 표본의 크기를 다시 선택하고 '표본추출(수동)'이나 '표본추출(자동)'을 클릭하고, '정규 모집단 만들기'부터 다시 하려면 상단의 '새로시작' 버튼을 클릭한다.

4) 2차시 결과 화면 [그림 III-19]



[그림 III-19] 2차시 결과 화면

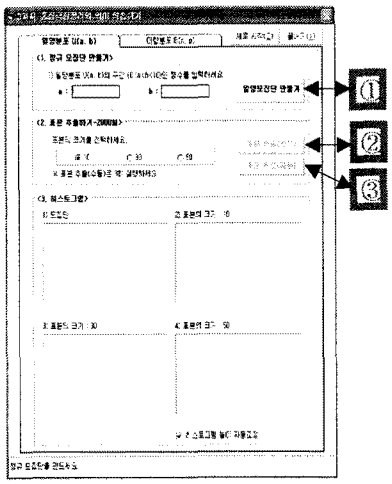
라. 중심극한정리의 의미 학습하기

1) 학습목표

일양분포 또는 베르누이분포를 따르는 모집단에서 표본을 추출하여 평균을 구해보고 구한 평균들의 히스토그램을 그려봄으로써, 정규분포를 따르지 않는 모집단도 그 표집분포가 정규분포를 근사적으로 따른다는 중심극한정리의 의미를 이해할 수 있다.

2) 일양분포

가) 초기 화면 [그림 III-20]



[그림 III-20] 3차시 일양분포 초기 화면

나) 수업 과정 ([그림 III-20] 참고)

(1) 일양모집단의 구간 'a'와 'b'를 입력하고 '일양 모집단 만들기' 버튼을 클릭한다. 일양모집단의 입력 범위는 0<a<b<10 인 정수로 설정되어 있으며, 잘못 입력 시 에러메시지가 pop-up 된다. 모집단의 히스토그램은 확률밀도함수의 그래프임을 유의하고 필요시 학생들에게 설명하도록 한다.

(2) 표본의 크기를 10으로 선택한 경우, '표본추출(수동)'을 클릭하여 추출된 표본들의 평균을 학생들이 직접 계산하여 입력하도록 한다. 평균을 잘 못 입력한 경우엔 정답을 알려준 후 다음

단계로 진행한다. 이 과정을 3회 반복한다.

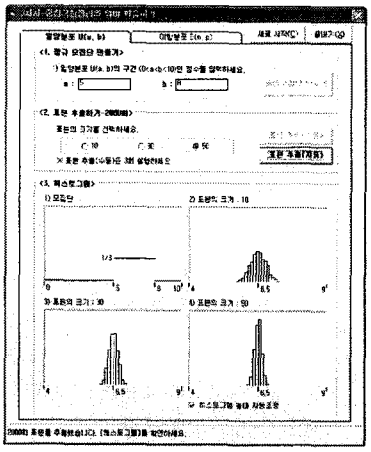
(3) '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하단의 상태표시영역에 작업진행표시 막대가 사라지면 표집분포의 히스토그램을 확인하고, 모집단의 히스토그램과 비교해본다.

(4) 표본의 크기가 30, 50인 경우, '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하단의 상태표시영역에 작업진행표시 막대가 사라지면 표집분포의 히스토그램을 확인하고, 모집단의 히스토그램과 비교해본다.

(5) 교사는 모집단이 정규분포를 따르지 않음에도 표집분포의 히스토그램이 정규분포 그래프와 비슷한 모양임을 강조하면서 중심극한정리의 의미를 설명한다.

(6) 표본추출을 다시 하려면 표본의 크기를 다시 선택하고 '표본추출(수동)'이나 '표본추출(자동)'을 클릭하고, '정규 모집단 만들기'부터 다시 하려면 상단의 '새로시작' 버튼을 클릭한다.

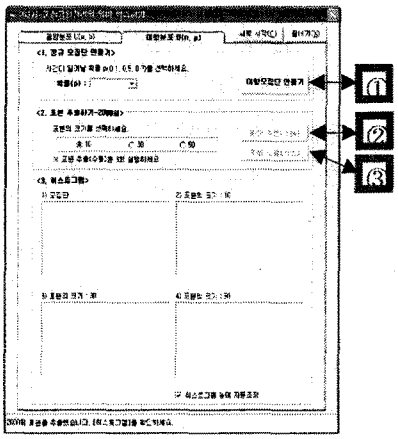
다) 결과 화면 (그림 III-21)



[그림 III-21] 3차시 일양분포 결과 화면

3) 이항분포

가) 초기 화면 [그림 III-22]



[그림 III-22] 3차시 이항분포 초기 화면

나) 수업 과정 ([그림 III-22] 참고)

(1) 이항모집단의 확률 'p'를 선택하고 '이항모집단 만들기' 버튼을 클릭한다. 모집단의 히스토그램을 확인한다. 이항분포는 사건이 일어날 확률이 'p'인 베르누이 시행을 n번 반복한 것이므로 모집단의 히스토그램이 두 개의 막대로 된 막대그래프 형태임을 설명한다.

(2) 표본의 크기를 10으로 선택한 경우, '표본추출(수동)'을 클릭하여 추출된 표본들의 평균을 학생들이 직접 계산하여 입력하도록 한다. 평균을 잘 못 입력한 경우엔 정답을 알려준 후 다음 단계로 진행한다. 이 과정을 3회 반복한다.

(3) '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하단의 상태표시영역에 작업진행표시 막대가 사라지면 표집분포의 히스토그램을 확인하고, 모집단의 히스토그램과 비교해본다.

교사는 표집분포의 히스토그램은 원래 막대그래프의 형태이어야 하나 이항분포의 연속성 보정을 사용하여 히스토그램처럼 그렸음을 설명한다. 또한 평균들의 히스토그램이므로 가로축의 좌표가 0 ~ 1의 값을 갖게 됨을 설명한다.

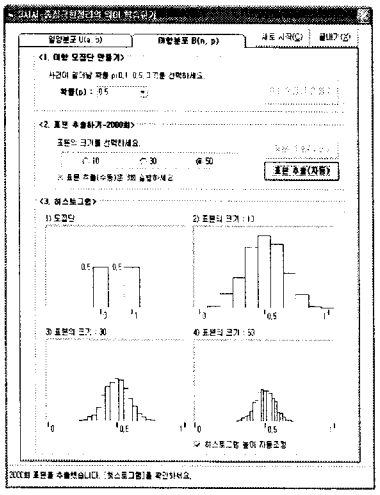
(4) 표본의 크기가 30, 50인 경우, '표본추출(자동)'을 클릭하여 자동추출을 실행시킨다. 하

단의 상태표시영역에 작업진행표시 막대가 사라지면 표집분포의 히스토그램을 확인하고, 모집단의 히스토그램과 비교해본다.

교사는 모집단이 정규분포를 따르지 않음에도 표집분포의 히스토그램이 정규분포 그래프와 비슷한 모양임을 강조하면서 중심극한정리의 의미를 설명한다.

(5) 표본추출을 다시 하려면 표본의 크기를 다시 선택하고 '표본추출(수동)'이나 '표본추출(자동)'을 클릭하고, '정규 모집단 만들기'부터 다시 하려면 상단의 '새로시작' 버튼을 클릭한다.

다) 결과 화면 : $p=0.5$ 인 경우 [그림 III-23]



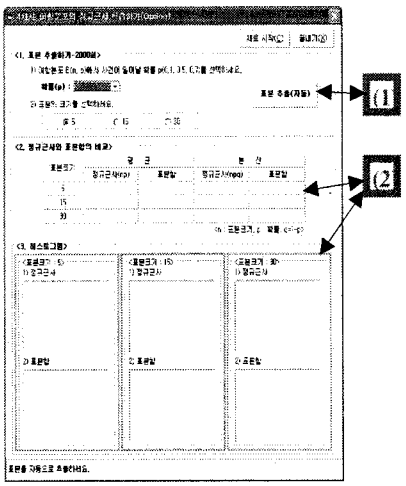
[그림 III-23] 3차시 이항분포 결과 화면

마. 이항분포의 정규근사 학습하기

1) 학습목표

이항분포를 따르는 모집단으로부터 추출한 표본들의 표본합과 정규근사 시의 평균, 분산, 히스토그램을 비교해봄으로써 표본의 크기가 충분히 클 때 이항분포가 정규분포로 근사함을 이해할 수 있다.

2) 4차시 초기 화면 [그림 III-24]



[그림 III-24] 4차시 초기 화면

3) 수업 과정 ([그림 III-24] 참고)

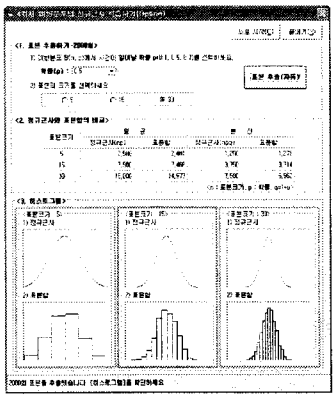
가) '확률(p)'와 '표본의 크기'를 선택한 후 '표본추출(자동)'을 클릭한다.

나) '2. 정규근사와 표본합의 비교'에서 평균과 분산을 비교해보고, 히스토그램도 비교해 본다. 교사는 p와 n의 값에 따라 정규근사 여부가 달라지는지 학생들이 확인해 보도록 한다.

다) 표본의 크기를 변경해 위의 과정을 반복한다.

라) 표본의 크기를 계속 변경하여 실행할 수 있으며, '확률'을 변경하고자 할 경우 '새로시작' 버튼을 클릭한다.

4) 4차시 결과 화면 : $p=0.5$ 인 경우



[그림 III-25] 4차시 결과 화면

IV. 결론 및 제언

본 연구에서는 학생들이 표집분포의 개념을 쉽게 이해할 수 있도록 '표집분포 시뮬레이션 (Sampling Distributions Simulation)'을 설계하고 Visual Basic 6.0을 사용하여 구현하였다.

본 연구는 선행연구에 비해, 우리나라 고등학교 통계 수업 현장에서 사용이 편리하도록 특별히 준비되었으며, 고교 통계교육과정을 반영한 여러 통계적 추론 학습과 병합하여 진행되도록 함으로서 통계적 추론 가운데 표집분포의 활용 모습이 잘 드러나도록 노력하였다.

본 연구를 통해 고등학교 통계교육에서 표집분포를 새롭게 인식하는 계기가 될 것을 기대하며 아울러 본 연구를 통해 제공[10]되는 소프트웨어가 많은 수업에서 적극 활용되어, 학생들이 통계를 기피하지 않고 오히려 통계에 대한 흥미와 관심이 증대되고, 나아가 사회생활을 할 때 통계에 대한 바른 인식을 갖고 좀 더 과학적인 태도로 살아갈 수 있는 토대가 형성되기 바란다.

참고문헌

강문봉, 송상헌, 이종희, 박선화 (2003). 수학과 ICT 활용 교수·학습 방법 개발. 대한수학교육학회 제24회 **추계학술대회 논문집**, 769-805.

강행고(1998). 제7차 수학과 교육과정 개정의 기본 방향. **수학교육 프로시딩, 7**, 7-19.

김동제 · 박용범(2001). 엑셀을 활용한 통계 수업의 흥미도 신장 방안. **학교수학, 3**(1), 109-129.

김민경(1998). 컴퓨터를 활용한 교수·학습의 구성주의적 접근. **교육학연구, 36**(2), 183-202.

나귀수(2000). 수학교육에서 컴퓨터 활용에 대한 소고. **학교수학, 2**(1), 97-110.

남주현(2007). **초·중등 통계교육을 위한 통계적 방법론에 대한 연구**. 이화여자대학교 대학원 박사학위논문.

노지희(2009). **미니탭을 활용한 고등학교 확률 및 통계교육** -7차개정교육과정 중심으로-. 한서대학교 교육대학원 석사학위논문

손미선(2003). **중등 수학 교육에서 컴퓨터의 활용에 관한 연구**. 서강대학교 교육대학원 석사학위논문.

신보미 · 이경화(2006). 컴퓨터 시뮬레이션을 통한 통계적 확률 지도에 대한 연구. **수학교육학연구, 16**(2), 139-155.

우정호(2000). 통계교육의 개선방향 탐색. **학교수학, 2**(1), 1-27.

우정호(2001). **학교수학의 교육적 기초(증보판)**. 서울: 서울대학교 출판부.

윤현진, 박선용, 김서령, 이영하 (2009). **수학과 교육내용 개선 방안 연구** -'이산수학', '확률과 통계' 영역을 중심으로-. 한국교육과정평가원. 연구보고 RRC 2009-3-3.

이경화(1996). **확률개념의 교수학적 변환에 관한 연구**. 서울대학교 대학원 박사학위논문.

이영하(2006). 통계적 개념의 발달과 통계교육. 서울 수학교육교원 1정연수 교재.

이영하(2009). 확률과 통계 영역의 지도. 광주 수학교육교원 1정연수 교재.

이외숙 외 (2002). 통계학 입문. 서울: 경문사.

이종영(1999). 컴퓨터 환경에서의 수학 학습-지도에 관한 교수학적 분석. 서울대학교 대학원 박사학위논문.

이진아(2007). 엑셀을 이용한 확률과 통계 교육에 관한 연구. 연세대학교 교육대학원 석사

10) 본 논문의 별첨 부록; 통계교육용 소프트웨어 <표집분포 시뮬레이션> 압축 파일은 대한수학교육학회 및 이화여대 수학교육과 홈페이지에 탑재되어 있으며 무료로 자유롭게 다운 받아 사용할 수 있음./source code는 이화여대 교육대학원 2010년 이은호의 석사논문의 부록에서 찾을 수 있음.

학위논문.

장형도(2006). **컴퓨터를 활용한 중등학교 확률 및 통계교육**. 연세대학교 교육대학원 석사학위논문.

Bakker, A. (2004). Design research in statistics education : on symbolizing and computer tools, Doctorial dissertations.

Batanero et al. (1994). Errors and difficulties in understanding elementary statistical concepts, International Journal of Mathematics Education in Science and Technology, 25(4), 527-547.

Behrens, J. T. (1997). Toward a theory and practice of using interactive graphics in statistics education. In J. B. Garfield & G. Burril (Eds.), Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 International Association for Statistics Education (IASE) roundtable, 111-122. Voorburg, The Netherlands: International Statistical Institute.

Cobb, G. W. & Moore, D. S. (1997). Mathematics, statistics and teaching. The American Methematical Monthly, Nov., 801-823.

Davenport, E. C. (1992). Creating data to explain statistical concepts: Seeing is believing. In Proceedings of the Section on Statistical Education of the American Statistical Association, 389-394.

delMas, R. C., Garfield, J., & Chance, B. (2004). Using Assessment to Study the Development of Students' Reasoning about Sampling Distributions. the American Educational Research Association.

Fisher, R. A. (1925). Statistical Methods for research workers. Oliver and Boyd of Edinburgh in Scotland.

Freudenthal, H. (1981) Major problems of mathematics education, Educational Studies in Mathematics, 12, p33-150

Glencross, M. J. (1988). A practical approach to the Central Limit Theorem. In R. Davidson & J. Swift (Eds.), Proceedings of the Second International Conference on Teaching Statistics, 91-95. Victoria, B.C.: Organizing Committee for the Second International Conference on Teaching Statistics.

Mendenhall, W., Wackeryly, D. D., & Scheaffer, R. L. (1990). Mathematical statistics with applications. Boston: PWS-KENT.

Moore, D. S. (1992). What is statistics?. In D. C. Hoaglin & D. S. Moore(Eds.), Perspectives on contemporary statistics (pp. 1-17). The Mathematical Association of America.

National Council of Teachers of Mathematics (1989). Curriculum and evaluation statndards for schoold mathematics. Reston, VA: National Council of Teachers of Mathematics. 구광조 외 역(1992). 수학교육과정과 평가의 새로운 방향, 서울: 경문사.

National Council of Teachers of Mathematics (2000). Principles and standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Rubin, A. & Bruce, B. (1991). Using computers to support students' understanding of statistical inference. New England Mathematical Journal.

Schwarz, C. J. & Sutherland, J. (1997). An on-line workshop using a simple capture-recapture experiment to illustrate the concepts of a sampling distribution. Journal of Statistics

Education, 5(1).

Simon, J. L. (1994). What some puzzling problems teach about the theory of simulation and the use of resampling. American Statistician, 48(4), 290-293.

Tall, D. (1991). 고등수학적 사고. 류희찬 외 역 (2003). 서울: 경문사.

The Design and Implementation to Teach Sampling Distributions with the Statistical Inferences

Lee, Young Ha (Ewha Womans University)
Lee, Eun Ho (The Graduate School of Education Ewha Womans University)

The purpose of the study is designing and implementing 'Sampling Distributions Simulation' to help students to understand concepts of sampling distributions. This computer simulation is developed to help students understand sampling distributions more easily.

'Sampling Distributions Simulation' consists of 4 sessions.

'The first session - Confidence level and confidence intervals - includes checking if the intended confidence level is actually achieved by the real relative frequency for the obtained sample confidence intervals containing population mean. This will give the students clearer idea about confidence level and confidence intervals in addition to the role of sampling distribution of the sample means among those.

'The second session - Sampling Distributions - helps understand sampling distribution of the sample means, through the simulation method to make comparison between the histogram of sampling distributions and that of the population

'The third session - The Central Limit Theorem - includes calculating the means of the samples taken from a population which follows a uniform distribution or follows a Bernoulli distribution and then making the histograms of those means. This will provides comprehension of the central limit theorem, which mentions about the sampling distribution of the sample means when the sample size is very large.

'The forth session - the normal approximation to the binomial distribution - helps understand the normal approximation to the binomial distribution as an alternative version of central limit theorem.

With the practical usage of the shareware 'Sampling Distributions Simulation', we expect students to have a new vision on the sampling distribution and to get more emphasis on it. With the sound understandings on the sampling distributions, more accurate and profound statistical inferences are expected. And the role of the sampling distribution in the inferences should be more deeply appreciated.

* **key words** : sampling distribution(표집분포), simulation method(모의실험), statistical inference(통계적 추론)

논문접수 : 2010. 7. 16
논문수정 : 2010. 9. 2
심사완료 : 2010. 9. 11