Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 359~369

Computer Simulation Program for Central Limit Theorem - Dynamic MS Excel Program -

Hyun Seok Choi[1] · Tae Yoon Kim[2]

Abstract

Central limit theorem is known as one of the most important limit theorem in statistics and probability. This paper provides a dynamic MS Excel program that demonstrates computer simulation of various types of central limit theorems. Our result will be of great use for better understanding of central limit theorems.

Keywords : Central Limit Theorem, Computer Simulation, Dynamic Excel Macro

1. 서론

통계학이나 확률론에서 중심극한정리는 실제 응용이나 이론 관점에서 가장 중요한 역할을 하고 있다고 할 수 있다. 통계학에서 키나 몸무게 같은 다양한 관찰현상들을 정규분포라고 가정하는 것이나 데이터가 많은 경우 표본평균 등의 분포가 정규분포를 따른다고 판단하는 것 등은 모두 중심극한정리 때문에 가능하다고 할 수 있다. 또한 확률론에서 다양한 통계량의 점근분포 도출 등이 가능한 이론적 배경도 역시 중심극한정리라고 할 수 있다. 따라서 중심극한정리를 좀 더 쉽게 이해할 수 있다면 통계학이나 확률론의 전반적 이해뿐만 아니라 실제 생활의 응용에도 많은 도움이 되리라 기대된다. 본 연구자들의 강의 경험에 의하면 많은 학생들이 중심극한정리를 까다로운 수학 정리의 일종으로 간주하거나 혹은 데이터들의 관찰수가 증가하면 데이터의 분포가 정규분포로 "변하는 현상"이라고 잘못 이해하는 경우가 종종 있다. 즉 중심극한정리의 기본 개념 『데이터들을 (혹은 요인들을) 많이 합하면 그 "합(sum)"의 분포가 종모양의 정규분포에 접근한다』는 사실을 제대로 이해하지 못하는 경우가 종종 있

1) First Author : Lecturer, Department of Statistics, Keimyung University, Daegu, 704-701, Korea
E-Mail : chsuk1@kmu.ac.kr
2) Professor, Department of Statistics, Keimyung University, Daegu, 704-701, Korea
E-Mail : tykim@kmu.ac.kr

다.

본 논문의 주된 목적은 중심극한정리의 기본 개념을 누구나 쉽게 이해할 수 있도록 도와주는 간단한 컴퓨터 실험 프로그램(동적 컴퓨터 프로그램)을 개발하는데 있다. 즉 중심극한정리의 동적 시뮬레이터의 개발이다. 이와 같은 동적 시뮬레이터 실험을 통한 중심극한이론의 이해라는 접근 방식은 확률론에서 알려진 다양한 특정 통계량 (예를 들어 U 통계량 등)의 점근 분포를 이해, 실험하는데도 손쉽게 확장될 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2절에서는 이제까지 연구된 중심극한정리의 구체적 내용들을 소개하고 이들에 대하여 개발된 컴퓨터 시뮬레이터를 이용한 실험의 과정 및 결과를 설명한다. 3절에서는 동적 시뮬레이터의 프로그램에 관한 기술적 설명을 포함하고 있고 4절에서는 결론을 맺고자 한다.

2. 중심극한정리

최초의 중심극한정리는 DeMoivre가 1773년경 X_i가 $p=1/2$인 베르누이 확률변수인 경우에 증명을 했으며 이후 다양한 경우로의 확장이 이루어졌다. 먼저 1812년에 Laplace가 임의의 p인 경우로 확장시켰다.

(중심극한정리 Ⅰ: DeMoivre-Laplace 극한정리)
S_n을 각 시행에서 성공할 확률이 p인 n번의 독립시행을 행할 때 발생하는 성공의 회수라 하면, 임의의 $a<b$에 대해, $n\to\infty$일 때

$$P\left\{a\le \frac{S_n-np}{\sqrt{np(1-p)}}\le b\right\}\to \Phi(b)-\Phi(a) \tag{2.1}$$

가 성립한다.

그 후 중심극한정리는 좀 더 일반적인 형태로 발전되었는데 동일하고 독립적인 분포를 따르는 확률변수들이 2차 적률을 갖는다는 조건하에서 아래와 같은 중심극한정리로 개선되었으며 이것이 일반적으로 기초 통계학이나 확률론에서 중심극한정리로써 소개되는 정리이다.

(중심극한정리 Ⅱ)
$X_1, X_2, \cdots, X_n$이 독립이고 각각 평균이 μ이고 분산이 σ^2인 동일한 분포를 따르는 확률변수 열 (sequence of random variables)이라 하자.
$\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$의 분포는 $n\to\infty$일 때, 표준정규분포를 따른다. 즉 $n\to\infty$일 때

$$P\left\{\frac{X_1+\cdots X_n-n\mu}{\sigma\sqrt{n}}\le a\right\}\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}e^{-x^2/2}dx \tag{2.2}$$

가 성립한다.

위의 정리 이후 동일한 분포라는 가정과 독립이라는 가정을 완화시켜주기 위한 중

심극한정리가 만들어 졌다.

(중심극한정리 Ⅲ)

$X_1, X_2, \cdots, X_n$을 각각 평균 $\mu_i = E[X_i]$와 분산 $\sigma_i^2 = Var(X_i)$인 독립 확률변수 열이라 하자. X_i들이 일양유계(uniformly bounded), 즉 모든 i에 대해 어떤 M이 존재하여 $|X_i| \le M$ 이고, $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$이면, $n \to \infty$일 때

$$P\left\{ \frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \le a \right\} \to \Phi(a) \tag{2.3}$$

가 성립한다.

그리고 독립성 조건을 완화시켜주는 중심극한정리는 다음과 같이 주어진다.

(중심극한정리 Ⅳ)

$X_1, X_2, \cdots, X_n$을 각각 평균 $\mu_i = E[X_i]$와 분산 $\sigma_i^2 = Var(X_i)$이며 약 의존적인 정상 확률변수(weakly dependent stationary random variable) 열이라 하자. $0 < c$에 대해 $Var(\sum_{i=1}^{n} X_i) = n\sigma^2 + n\sum_{i \ne j}^{n} Cov(X_i, X_j) \sim nc$이면, $n \to \infty$일 때

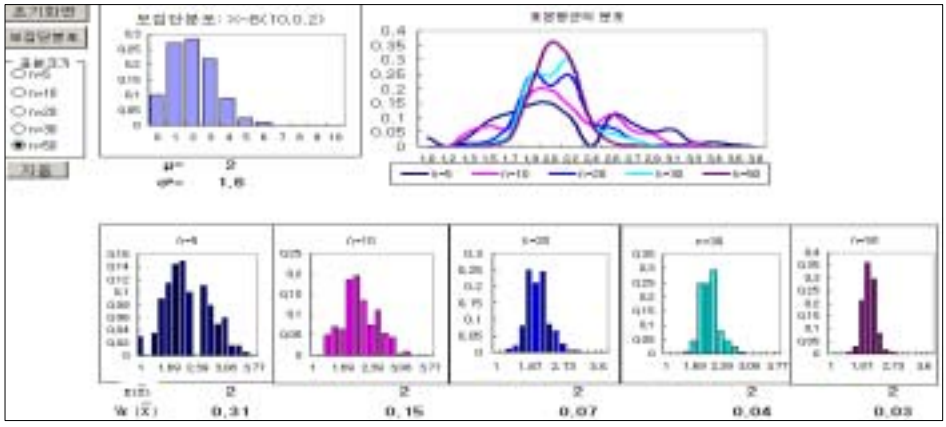
$$P\left\{ \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{Var(\sum_{i=1}^{n} X_i)}} \right\} \to \Phi(a) \tag{2.4}$$

가 성립한다.

물론 이외에도 다양한 중심극한정리들이 존재하나(Billingsley, 1963, 1986) 본 논문에서는 논의의 단순성을 위해 위의 네 개의 중심극한정리들에 대해서 컴퓨터 실험을 고려한다.

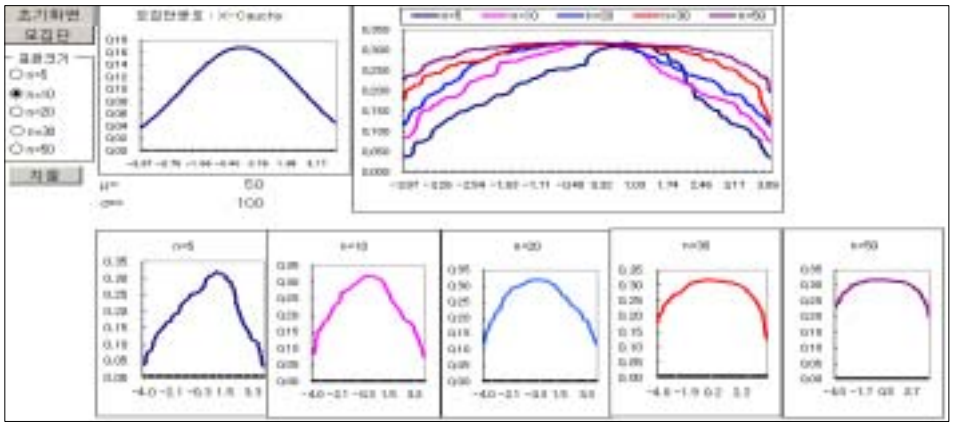
중심극한정리에 대한 실험프로그램은 개인용 PC에 대부분 설치되어있고, 업무용으로 많이 사용되어 일반인에게 익숙한 Excel을 이용하여 개발하였다. 프로그램의 초기화면은 <그림 1>과 같다. <그림 1>에는 다양한 분포와 의존구조들이 주어져 있으며 해당 모집단의 분포를 마우스로 클릭하면 각각의 모집단분포로 이동하게 된다. 예를 들어 이항분포를 클릭하면 <그림 2>가 나타나서 중심극한정리 I이 뜻하는 바를 실험하게 된다. 표본 크기의 옵션단추를 클릭함으로써 표본크기 증가에 따라 표본평균의 분포가 정규분포에 근사됨을 동적으로 확인할 수 있게 된다. 참고로 본 논문의 모든 실험 프로그램에서는 $\overline{X_n} = S_n / n$의 점근 정규성(asymptotic normality)에 초점을 맞추어서 n이 클 경우 $\overline{X_n} \approx N(\mu, \sigma^2/n)$이 됨을 확인하고자 하였다.

<그림 1> 초기화면

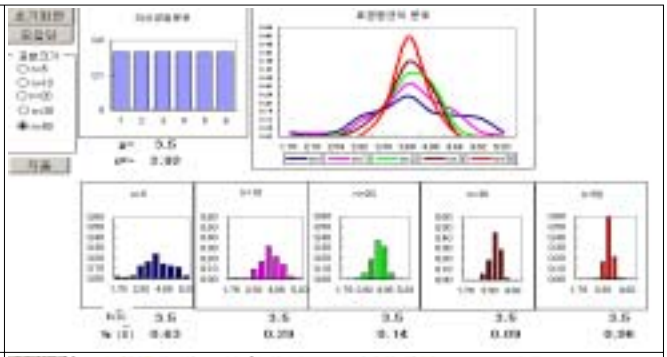
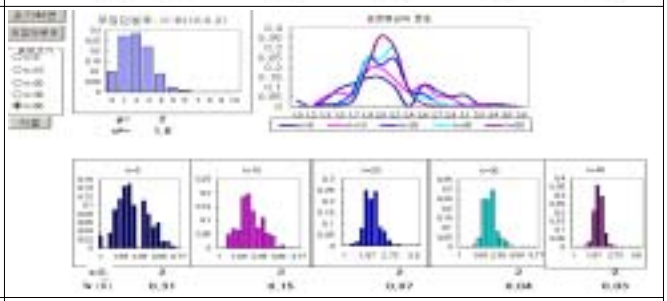
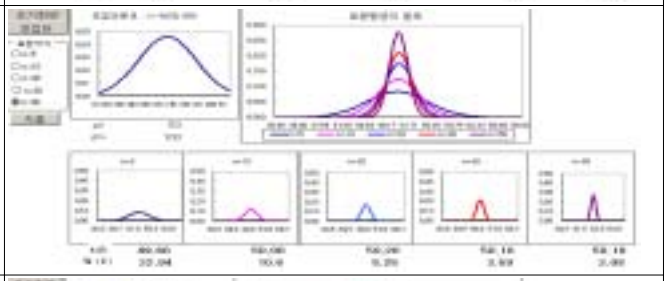
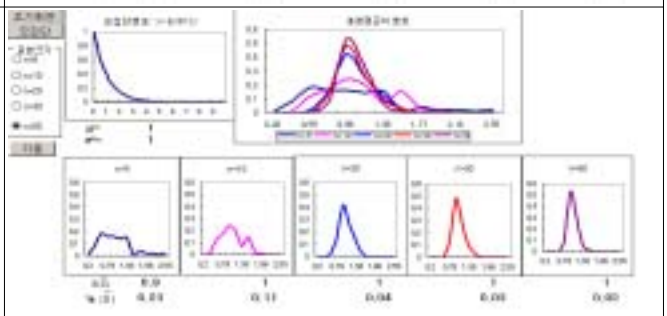
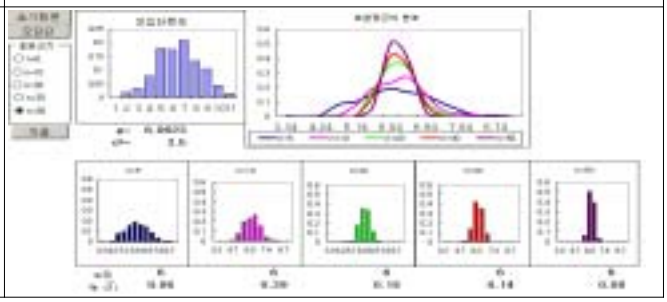


<그림 2> 이항분포의 중심극한정리 화면

중심극한정리 Ⅱ에 대한 실험은 <그림 4>에 주어져 있다. 이 경우 분포의 2차 적률의 존재여부가 중심극한정리에 어떠한 영향을 주는지 알아보기 위해서 2차 적률이 존재하지 않는 것으로 알려진 코쉬(Cauchy)분포를 실험해보면 그 결과는 <그림 3>에 주어진다. <그림 3>에서 보면 코쉬분포의 경우 n이 증가함에 그 모양이 둥그런 모양(round shape)으로 접근함을 알 수 있으며 정규분포 모양과는 다른 모양이 된다.

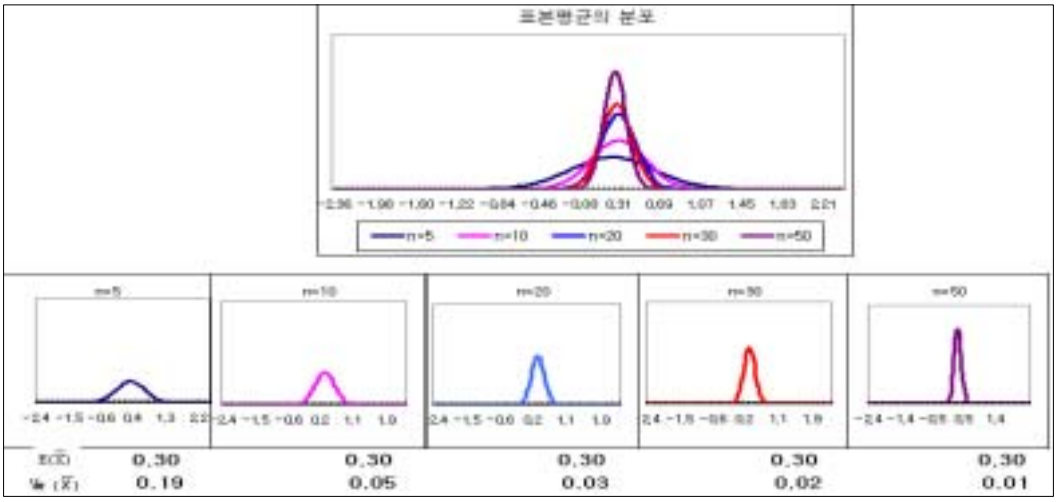


<그림 3> Cauchy분포의 중심극한정리 실험

이산 균일 분포	$\overline{X}$가 평균이 $\frac{a+b}{2}$, 분산이 $\frac{(b-a)^2}{12}$인 균일분포에서의 표본평균이면, $\overline{X}$는 $N\left(\frac{a+b}{2}, \frac{(b-a)^2}{12n}\right)$에 근사한다는 것을 확인할 수 있다.	
이항 분포	$\overline{X}$가 평균이 np, 분산이 npq인 이항분포에서의 표본평균이면, $\overline{X}$는 $N(np, \frac{npq}{n})$에 근사한다는 것을 확인할 수 있다.	
정규 분포	$\overline{X}$가 평균이 μ, 분산이 σ^2인 정규분포에서의 표본평균이면, $\overline{X}$는 $N(\mu, \frac{\sigma^2}{n})$을 따르는 것을 확인할 수 있다.	
지수 분포	$\overline{X}$가 평균이 $1/\lambda$, 분산이 $1/\lambda^2$인 지수분포에서의 표본평균이면, $\overline{X}$는 $N(1/\lambda, 1/n\lambda^2)$에 근사한다는 것을 확인할 수 있다.	
삼각형 분포	$\overline{X}$가 평균이 μ, 분산이 σ^2인 임의의 분포에서의 표본평균이면, $\overline{X}$는 $N(\mu, \frac{\sigma^2}{n})$에 근사한다는 것을 확인할 수 있다.	

<그림 4> 중심극한정리 Ⅰ과 Ⅱ

중심극한정리 Ⅲ에 대한 실험은 <그림 5>에 나타나 있는데 홀수 번째 확률변수는 정규분포, 짝수 번째는 균일분포를 하는 경우를 고려하였으며 이 경우 정규분포로의 접근 모습이 <그림 2>의 모든 변수들이 동일한 분포를 하는 경우와 크게 다르지 않음을 알 수 있다.

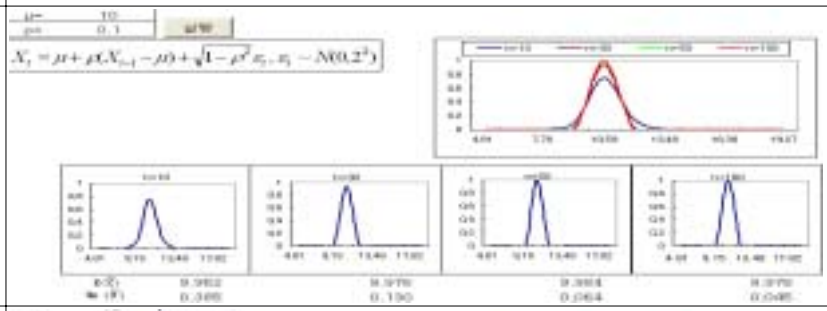
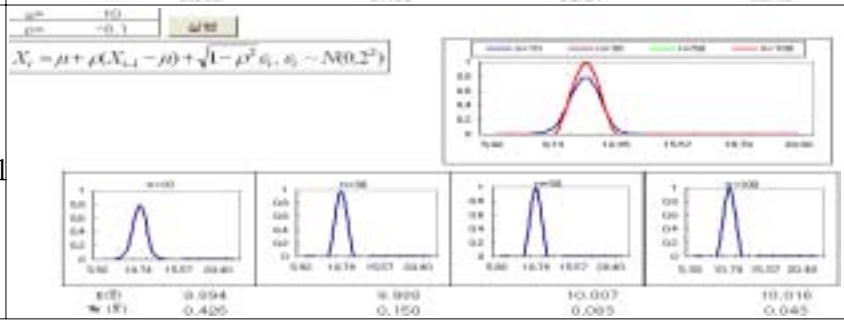


<그림 5> $X_i = N(0,1)$과 $X_{i+1} = U(0,1)$에 대한 중심극한정리

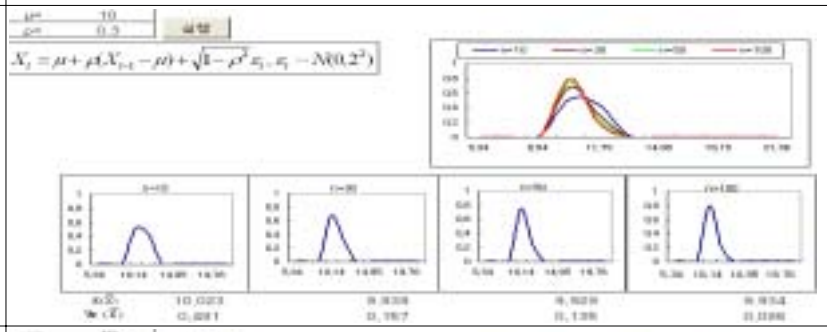
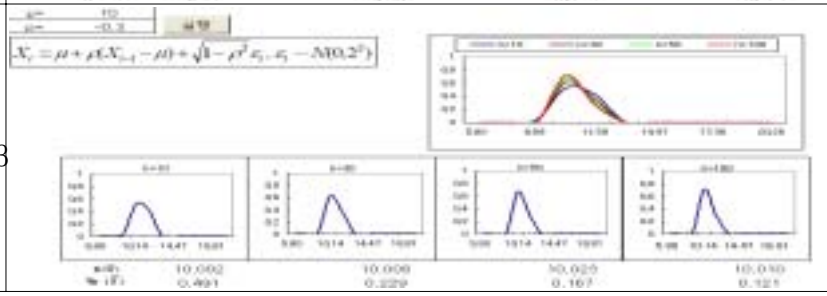
의존 변수들의 중심극한정리 Ⅳ에 대한 실험은 <그림 6>과 <그림 7>에 주어져 있다. 이를 통해 쉽게 확인할 수 있는 바는 ρ가 양수인 경우와 ρ가 음수인 경우의 $\overline{X}$의 분산이 조금씩 다르다는 점이다. 이것이 바로 의존성의 효과라고 할 수 있는데 그 이유는

$$Var(\sum_{i=1}^{n} X_i) = nVar(X_i) + n\sum_{i \neq j}^{n} Cov(X_i, X_j) \tag{2.5}$$

에서 ρ의 크기에 따라 $Cov(X_i, X_j)$의 값이 변하게 되어 $Var(\sum_{i=1}^{n} X_i)$값에 영향을 미치기 때문이다. 즉 $|\rho|$의 값이 커짐에 따라 $Var(\sum_{i=1}^{n} X_i)$값이 증가하는 것으로 알려져 있으며 여기서 기억해야 될 점은 X_i의 분포는 ρ와 관계없이 일정하다는 점이다.

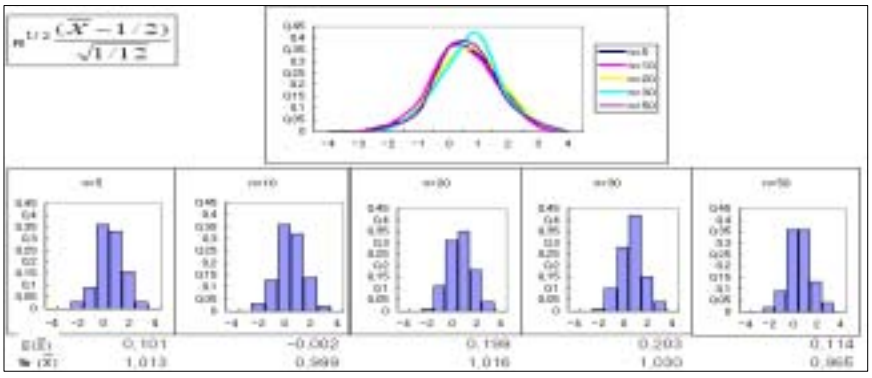
	$X_i = \mu + \rho(X_{i-1} - \mu) + \sqrt{1-\rho^2}\ \epsilon_i,\ \ \epsilon_i \sim N(0, 2^2)$
$\mu = 10$ $\rho = 0.1$	
$\mu = 10$ $\rho = -0.1$	

<그림 6> $X_i = \mu + \rho\,(X_{i-1} - \mu) + \sqrt{1-\rho^2}\epsilon_i,\ \ \epsilon_i \sim N(0, 2^2)$ 일 때 중심극한정리

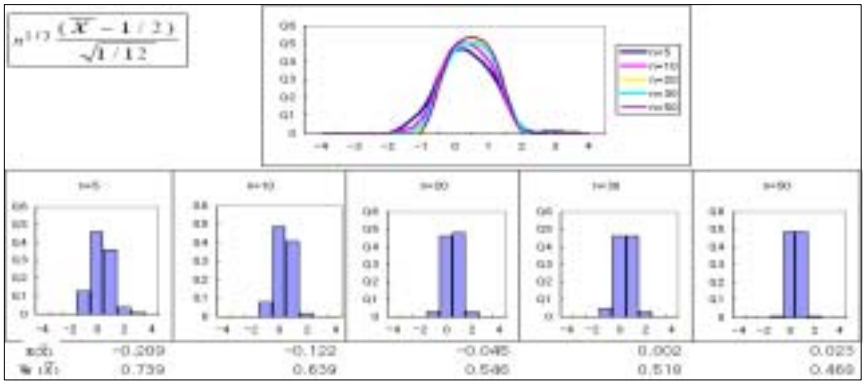
	$X_i = \mu + \rho(X_{i-1} - \mu) + \sqrt{1-\rho^2}\ \epsilon_i,\ \ \epsilon_1 \sim N(0, 2^2)$
$\mu = 10$ $\rho = 0.3$	
$\mu = 10$ $\rho = -0.3$	

<그림 7> $X_i = \mu + \rho\,(X_{i-1} - \mu) + \sqrt{1-\rho^2}\epsilon_i,\ \ \epsilon_i \sim N(0, 2^2)$ 일 때 중심극한정리

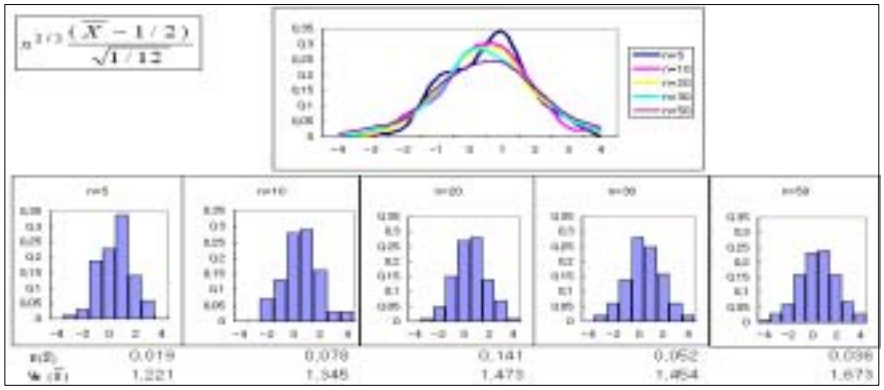
마지막으로 중심극한정리의 수렴속도에 대한 컴퓨터 실험 결과를 고려해보자. 중심극한정리의 수렴속도는 $n^{1/2}$이라고 알려져 있다. 즉 표본평균의 모평균으로 수렴속도는 표본평균의 표준편차의 차수(order) $n^{1/2}$이며 이에 대한 컴퓨터 실험결과는 <그림 8> ~ <그림10>에 주어져 있다. <그림 8>에서는 본래의 수렴속도인 $n^{1/2}$로 scale된 경우를, <그림 9>와 <그림 10>에서는 $n^{1/3}$, $n^{2/3}$로 scale된 경우를 각각 실험하였다. 이론적으로 기대되는 바는 <그림 9>에서는 표본평균이 모평균에 더욱 가깝게 되고 <그림 10>에서는 표본평균이 모평균으로부터 더욱 멀어지게 되어 결과적으로 두 경우 모두 표준 정규분포의 모양을 잃게 될 것이라는 것이다. 실제로 컴퓨터 실험 결과인 <그림 9>와 <그림 10>은 이와 같은 사실을 확인해 주고 있다.



<그림 8> $n^{1/2}\frac{(\bar{X}-1/2)}{\sqrt{1/12}}$의 중심극한정리



<그림 9> $n^{1/3}\frac{(\bar{X}-1/2)}{\sqrt{1/12}}$의 중심극한정리



<그림 10> $n^{2/3}\frac{(\overline{X}-1/2)}{\sqrt{1/12}}$의 중심극한정리

3. 프로그램 설명

중심극한정리 개발에 사용한 도구는 엑셀을 기반으로 하여 프로그램 제어와 함수 사용, 설명, 그래프 등을 위하여 양식도구, 매크로, VBA(Visual Basic for Application)를 사용하였다. 양식도구는 Dialog Sheet상에서 대화상자를 사용자가 직접 작성할 때 사용하는 것으로 프로그램에서는 명령단추(Command Button), 옵션단추(Option Button), 그룹상자(Group Box)를 사용하였다. 매크로는 이용자의 작업 처리를 코드로 기록해 두었다가 나중에 이 코드로 작업을 자동으로 수행하기 위한 명령으로 엑셀매크로의 경우 반복적으로 실행되는 명령어를 모아서 한 번에 실행하는 것 외에 비주얼 베이직이라는 프로그램언어를 내장하여 단순한 명령어 나열이 아니라 명령어를 이용한 프로그래밍이 가능하도록 되어 있다. VBA Project의 모듈 창에는 다음과 같이 코드를 작성하였다.

첫째, Sub 문을 사용하여 일반프로시저로 작성하였다.

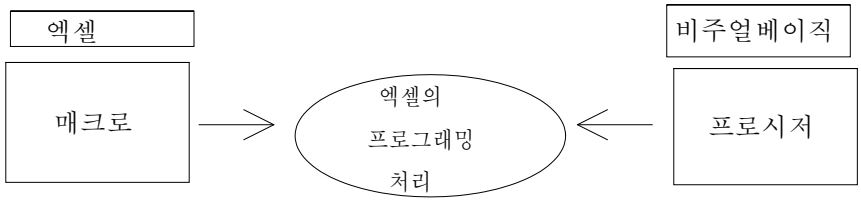
둘째, 셀을 지정하여 입력받은 값을 수식을 이용하여 기록하는 프로시저를 작성하였다.

셋째, 엑셀 자체에서 제공되는 분석기능과 VBA 등으로 작성된 프로그램을 연결하였다(Jacobson, 2002).

넷째, 설명, 수식, 그래프 등은 Rectangle, Object, Chartobjects 등으로 작성하여 ActiveSheet를 사용하여 활성화하였다.

다섯째, 명령단추를 사용하여 단추를 누르면 바로 매크로가 실행되게 하였다.

위의 절차를 그림으로 나타내면 아래와 같다.



<그림 11> 엑셀 매크로와 비주얼베이직의 프로시저

본 프로그램은 <표 1>의 경우에 대하여 프로그램을 실행할 수 있도록 구성되어 있다.

<표 1> 중심극한정리에 사용된 함수

독립인 표본	분포형태	난수 발생 함수식
	이산균일분포	=INT(a+(b-a+1)*RAND()) (a=최소값, b=최대값)
	이항분포	=CRITBINOM(x,p,RAND()) (x=성공회수, p=성공확률)
	정규분포	=NORMINV(RAND(),μ,σ) (μ=평균, σ=표준편차)
	지수분포	=(-1/λ)*LN(RAND()) (λ=평균)
	삼각형형태	=a+(b-a)*(RAND()+RAND())/2 (a=최소값, b=최대값)
시계열 데이터	$X_i = \mu + \rho(X_{i-1} - \mu) + \sqrt{1-\rho^2}\,\varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_i^2)$	
	$X_i = \mu + \rho(X_{i-1} - \mu) + \sqrt{1-\rho^2}\,\varepsilon_i, \quad \varepsilon_i \sim U(0,1)$	

4. 결론

본 프로그램은 모집단의 분포의 형태와 표본의 크기에 따라 GUI(Graphic User Interface)환경에서 버튼을 누름으로써 실행이 되도록 하였다. 모집단의 분포와 무관하게 표본의 수가 증가함에 따라 표본평균의 분포가 정규분포에 가까워지는 것을 확인할 수 있다. 특히 그래프의 변화 상태를 마우스의 클릭으로 동적으로 비교할 수 있다는 데 그 주된 특징이 있다. 향후 확률론에서 알려진 U 통계량 등의 다양한 특정 통계량 점근 분포에도 적용하면 이론을 효과적으로 이해하는데 기여할 수 있을 것이다.

참고문헌

1. 김태윤, 신기동 (1996). 중심극한정리에 대한 연구, *수리과학논집*, 제16집, 제1호, 89-95.
2. Billingsley, P. (1963). *Convergence of Probability Measures*, John Wiley & Sons. Wiley, New York.
3. Billingsley, P. (1986). *Probability and Measure*, John Wiley & Sons. Wiley, New York.
4. Jacobson, R. (2002). *Microsoft Excel 2002 Visual Basic Step by Step*. Redmond, W.A.: Microsoft Press.

[2005년 2월 접수, 2005년 5월 채택]