The Korean Communications
in Statistics Vol. 6, No. 3, 1999
pp. 821-829

Approximation of Binomial Distribution via Dynamic Graphics

Songyong Sim[1] Kee-Won Lee[2]

Abstract

In This paper, we calculate the probabilities of binomial and Poisson distributions when n or μ is large. Based on this calculation, we consider the normal approximation to the binomial and binomial approximation to Poisson. We implement this approximation via CGI and dynamic graphs. These implementation are made available through the internet.

1. 서론

인터넷이 보편화되면서 많은 분야에서 인터넷을 이용하고 있다(심송용 (1997)). 이런 현상은 통계학 교육에서도 예외가 아니어서 여러 종류의 모의실험이나 통계이론 등을 인터넷을 통하여 제공하고 그 전달 방법이 멀티미디어적이어서 사용자가 이론을 이해하는데 도움을 준다. 이런 연구는 조신섭 등 (1997), 한경수 등 (1998), 강희모, 심송용 (1998), 안기수와 허문열 (1998), 이우리와 최현집(1998), Yilmaz(1996) 등의 연구 결과에 의해 잘 알려져 있다. 본 논문은 이러한 연구의 연장선 상에서 이항분포의 정규 근사를 JAVA 애플릿을 이용한 동적 그래프로 구현하여 인터넷을 통하여 강의실이나 집에서 스스로 공부할 수 있게 하였다. 특히 애플릿이나 CGI를 이용하는 경우에는 사용자가 Netscape 등 브라우저만 있으면 특정 프로그램을 받아와야 하거나 설치할 필요가 없이 바로 사용할 수 있는 장점이 있다.

이항분포 $B(n,p)$를 따르는 확률변수 X의 분포는 n이 커지면 기대값과 분산이 각각 np, $np(1-p)$인 정규분포 $N(np, np(1-p))$로 근사될 수 있다는 것과 이항분포는 또 포아송 분포로 근사할 수 있다는 것은 잘 알려진 사실이다. 이에 대한 이론적인 증명은 적률함수나 특성함수를 이용하여 하는 것이 일반적이다. 그러나 이러한 근사는 n이 커질 때의 근사이고 n이 조금만 커져도 이항분포나 포아송 분포의 확률을 계산하는 것은 단순한 작업이 아니다. 실제로 초기하분포나 포아송 분포에 대한 확률의 계산값은 책으로도 출판되어 있을 정도이다(Liberman and Owen(1961), GE(1962)). 본 연구에서는 상당히 모수가 큰 경우에까지 이러한 확률을 계산하고 이 계산의 결과를 동적 그림과 CGI(Common Gateway Interface)를 이용하여 인터넷으로 공개하였다. 여기에서 얻을 수 있는 값은 포아송분포의 경우 GE(1962)가 제공하는 값보다도 많다.

이항분포의 정규근사를 JAVA 애플릿으로 구현한 장소는 인터넷에서 찾기가 어렵지 않은데 모

1) Dept. of Statistics, Hallym Univ., 1 Okchun Chunchon, Korea 200-702. This research was supported by the 1999 Hallym University Research Fund.
2) Dept. of Statistics, Hallym Univ. 1 Ockchun Chunchun, Korea 200-702.

두 n이 조금만 커도 정규근사를 그림으로 제대로 구현하지 못하였다. 예를 들면 외국 사이트로는

http://www.stat.wvu.edu/~hxue/normalapprox.htm
http://stat-www.berkeley.edu/users/stark/Java/BinHist.html
http://stad.dsl.nl/~berriel/ctljava.html
http://www.cs.uni.edu/~campbell/stat/prob8.html
http://arch.econ.hku.hk/stat/west/javahtml/binom_demo.html

등 아주 많지만 실제로는 서너개 정도의(위의 1,2,3 번째) 서로 다른 애플릿이 있고 국내에는 위의 세번째부터 다섯번째와 같은 애플릿을 가지고 있는

http://stat.yonsei.ac.kr/applet/binomial.htm

등이 있다. 이런 애플릿의 한계는 대개 n이 100 내외이다. 본 고에서는 위의 애플릿 보다 훨씬 더 큰 n에 대해서도 작동할 수 있는 애플릿을 설계하였다. 또

http://www.stat.sc.edu/~west/applets/binomialdemo.html

에서는 n이 큰 경우의 계산은 가능하나 포아송 분포나 정규분포와 수렴성 비교는 하지 않고 있다.

2. 이항분포의 확률 계산

이항분포 $B(n, p)$를 따르는 확률변수 X의 확률밀도함수 $f(x; n, p)$는

$$\Pr[X=x]=f(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x} \qquad x=0,1,2,\ldots,n \tag{1}$$

인데 이 확률을 계산하려면 $\binom{n}{x}$의 값이 n이나 x의 값이 조금만 커져도 오버플로우(overflow)가 발생하기 때문에 이 식을 바로 사용할 수가 없다. 또한 p^x와 $(1-p)^{n-x}$의 값은 아주 작은 값이 되기 때문에 언더플로우(underflow)가 발생하기 쉽다. 확률 계산에서 이 둘을 피해 가기 위해서 다음과 같이 이 확률을 계산하였다.

1.먼저 $x>n/2$인 경우와 그렇지 않은 경우로 나누어서 $[y]$는 y를 넘지 않는 가장 큰 정수 값을 취하는 함수라 할 때 $x>n/2$ 이면

$$\frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\,p^x(1-p)^{n-x} \quad x=[n/2],[n/2]+1,\ldots,n \tag{2}$$

을 계산하고 그렇지 않으면

$$\frac{n(n-1)(n-2)\cdots(x+1)}{(n-x)!}\,p^x(1-p)^{n-x} \quad x=0,1,2,\ldots,[n/2] \tag{3}$$

를 계산하도록 하였다. 이렇게 계산함으로써 불필요한 계승(factorial) 계산을 줄일 수 있다.

2. 위의 식 (2)와 (3)도 이 식들에 표현된 순서대로 계산하면 여전히 오버플로우 와 언더플로우가 생기므로 계산의 순서는 다음과 같은 방법으로 하였다.
   즉 $x>n/2$인 경우에는

$$p^{2x-n}\prod_{i=0}^{n-x-1}(n-i)*p/(x-i)*(1-p) \tag{4}$$

로 계산하고 그렇지 않은 경우에는

$$(1-p)^{n-2x}\prod_{i=0}^{x-1}(n-i)*p/(n-x-i)*(1-p) \tag{5}$$

로 계산하였다.

3. 이 계산을 $x=[np]$에서 시작하여 $x=[np]\pm i$일 때 다음 계산을 하는 방법으로 하여 $i=i_0$일 때의 확률이 0으로 계산되면 $i>i_0$인 경우의 확률을 0으로 설정하였다. 이 방법으로 불필요한 확률 계산을 많이 줄일 수 있게 된다

   참고: 이렇게 하더라도 n이 큰 경우에는 계산 과정에서 지나치게 큰 수가 나오게 되는데 중간곱의 계산이 큰 경우는 p나 $(1-p)$를 필요한 만큼 곱하여 오버플로우가 나는 것을 방지하였다.

위의 알고리즘을 JAVA로 구현하여 그 계산 결과를 확인하였더니 그 결과 n의 값이 p의 값을 0.1, 0.5 및 0.9 일 때 n을 변화시키면서 이 알고리즘을 테스트하였는데 3000 내외일 때까지 아무 문제없이 계산되었다. 참고로 MINITAB version 10.2와 12.22에서 $n=500$일 때도 이항분포 확률은 계산되지 않았다. 한편 n이 상당히 커지더라도 (예: $n=100,100$) 올바른 계산 결과를 보여 주었다.

3. 포아송 확률의 계산

MINITAB의 경우 포아송 분포의 확률은 기대값 λ가 50을 넘으면 계산하지 못하는데 우리는 10000 정도까지는 계산할 수 있는 알고리즘을 사용하였다.먼저 $\exp\{-\lambda\}$를 계산하여 이 값이 컴퓨터 상의 0이면 λ를 반으로 취하여 다시 같은 값을 계산하는 과정을 반복하여 0이 아닌 값이 나올 때까지 반복한다. 이 계산을 한 마지막 값은

$$\prod_{i=1}^{x} \frac{\lambda}{i} \tag{6}$$

를 계산하는 중간에 1보다 커질 때마다 곱하여 주어 포아송 확률 계산에서 지나치게 큰 값과 작은 값이 나타나는 현상을 상쇄하였다.

4. 애플릿의 실행

애플릿에서는 축과 축에 해당하는 숫자 등을 검은색으로 그린 후에 정규분포 $N(np, np(1-p))$의 확률밀도함수를 붉은색으로 그리도록 하였다. 이 그림 위에 이항분포의 히스토그램은 파란색으로, 포아송 분포의 히스토그램은 녹색으로 그렸다. 이항분포 $B(n,p)$를 따르는 확률변수 X의 확률밀도함수는 평균이 np이고 분산이 $np(1-p)$인 정규분포에 근사하므로 확률 기둥그림과 정규분포의 확률밀도함수를 모든구간에서 다 그리지 않고

$$x \in (np-4\sqrt{np(1-p)}, np+4\sqrt{np(1-p)})$$

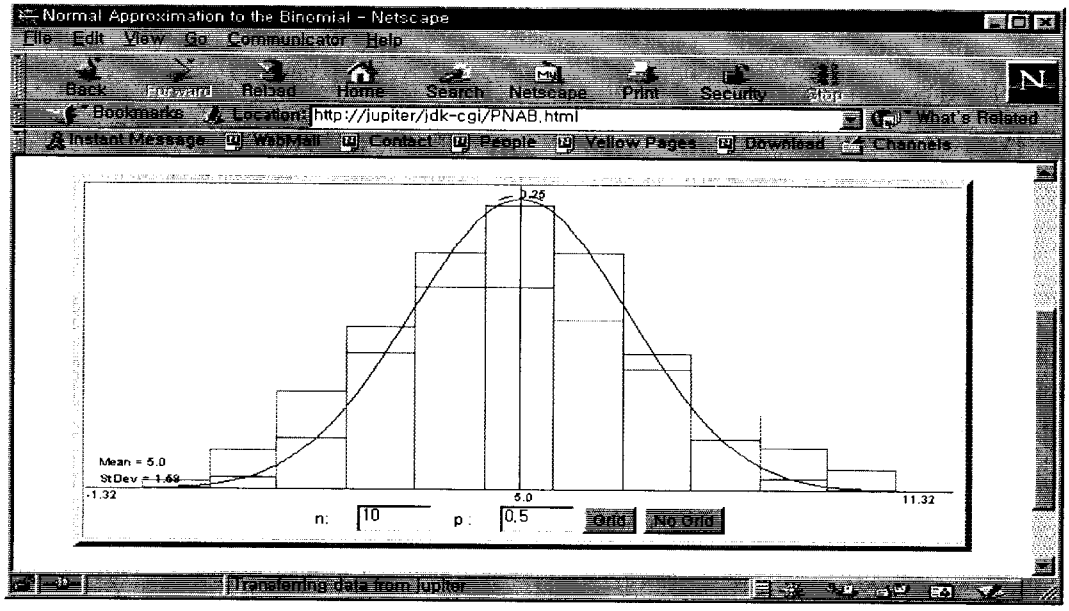


그림 1 : 초기 화면

인 x에 대해서만 그림을 그리도록 하였다. 이렇게 함으로서 정규분포의 확률밀도함수가

그림으로 의미가 있는 부분은 모두 그릴 수 있다. n을 불필요하게 크게 하는 것을 방지하기 위해 n이 5000보다 큰 경우에는 실행하지 않도록 하였다. n이 커지면 계산이 너무 많아져서 사용자가 사용하는 컴퓨터와 브라우저에 따라 다르긴 하지만 속도가 크게 저하되는 현상을 보일 뿐 아니라 계산 결과도 오차 범위를 벗어나기 때문이다. n이 취할 수 있는 가장 작은 값은 1이고 자연수가 아닌 n의 값에 대해서는 경고 메세지를 인쇄하도록 설계되었다. p가 취할 수 있는 값은 0과 1사이(경계점 불포함)인 실수인데 이 범위 밖의 p값을 사용자가 입력할 경우 역시 경고 메세지를 인쇄하도록 설계되었다. 또 포아송 분포의 확률을 계산할 때는 $\lambda = np$로 얻은 포아송 분포의 확률을 계산하였다. 포아송 확률 히스토그램은 녹색으로 그리도록 하여 흐리게 보이도록 하였다. 애플릿의 초기 값은 $n=10$, $p=0.5$로 하여 대칭이면서 정규분포와 비슷한 확률 도수분포 그림을 그리도록 하였다.
이 애플릿은

http://jupiter.hallym.ac.kr/jdk-cgi/PNAB.html

을 방문하면 누구나 얻을 수 있다. 이 애플릿은 JDK1.2로 작성된 것인데 JDK 1.1 및 그 후속 버전을 지원하는 브라우저로 볼 수 있다. 이런 브라우저로는 MS Explorer 4.0이후 버전, Netscape 4.5이후 버전 등이 있다. 이 장소를 방문하면 초기 화면은 그림 1과 같이 $n=10$이고 $p=0.5$인 경우가 뜨는데 사용자가 n값과 p값을 원하는 값으로 설정하고 이 그림에 보이는 "Grid" 버튼이나 "No Grid"버튼을 눌러

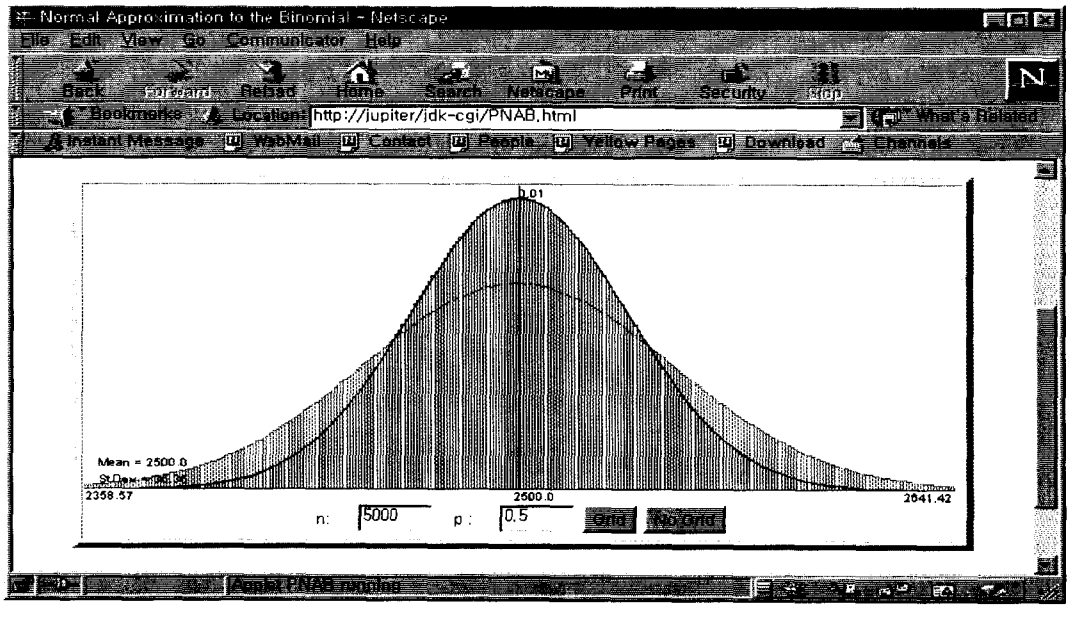


그림 2 : n이 크고 p가 0.5에 가까울 때: 이항분포(실선 기둥)와 정규분포는 모양이 거의 같으나 포아송 분포(흐린 기둥)와 이항분포는 모양이 많이 다름.

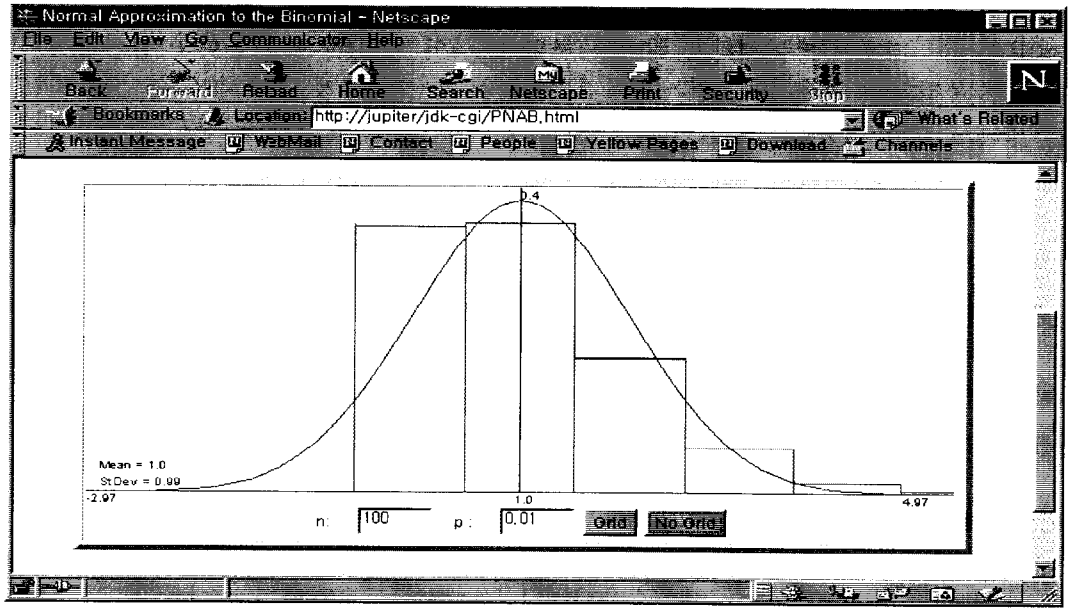


그림 3 : n이 크고 p가 0에 가까운 경우: 이항분포(실선의 기둥)와 포아송 분포(점선의 사각형)가 같아서 구별이 잘 안됨.

주면 새로 설정된 값에 따라서 그림을 다시 그린다. 이 애플릿에서 "Grid" 버튼을 사용하면 가로와 세로로 노란색의 보조선을 그어 준다. 초기 화면에서 보는 것은 이항분포는 정규분포에 근사하나 포아송 분포와 이항분포는 그렇지 않다는 것이다. 히스토그램의 기둥수가 많아지는 경우에 (주로 n이 아주 크고 p가 0.5 근처 일 때)는 각 기둥의 폭이 달라 보이는 경우가 있는데 이는 컴퓨터 그래픽이 그림을 픽셀(pixel) 단위로 하고 픽셀의 값은 음이 아닌 정수 밖에 사용할 수 없는 한계 때문인데 이 현상은 디지털 화면을 사용하는 한 피할 수 없다. 그림 2가 그런 경우의 예이다. 그림 2에서는 n이 적당히 크므로 이항분포와 정규분포의 근사는 거의 완벽하게 이루어지나 이항분포와 포아송 분포는 모양이 많이 다르다는 것을 알 수 있다. 이는 이항분포의 포아송 근사는 n이 크고 p가 작아서 np의 값이 상수로 수렴할 때 적용된다는 사실을 뒷받침한다(Johnson and Kotz(1969)). 이는 그림 3에서 n이 크고 p가 0에 가까운 값을 취한 경우, 그림의 이항분포가 정규분포에 근사하는 것보다는 포아송 분포에 훨씬 더 근사함을 그림으로 볼 수 있다.

5. 이항 확률과 포아송 확률 CGI

앞 절에서 본 애플릿을 구현하기 위해 이미 이항분포의 확률을 구하였으므로 이 절에서는 특정한 이항분포의 확률이 필요한 경우에는 숫자로 직접 그 확률을 확인할 수 있는 CGI를 구현하였다. 웹 브라우저 사용자가 확률 계산이 필요한 이항분포 $B(n,p)$의 모수 n과 p 및 확률계산이 필요한 확률변수 $X \sim B(n,p)$의 범위의 좌우 경계점 $a \leq b$를 입력하면 확률 계산은 CGI를 통

해 웹 서버에서 직접 하도록 하고 그 결과를 브라우저를 통하여 사용자에게 보여 주도록 하였다. 계산 결과는 사용자가 준 범위 안의 각각의 $X = x(a \leq x \leq b)$값에 대한 확률과 그 누적확률 및

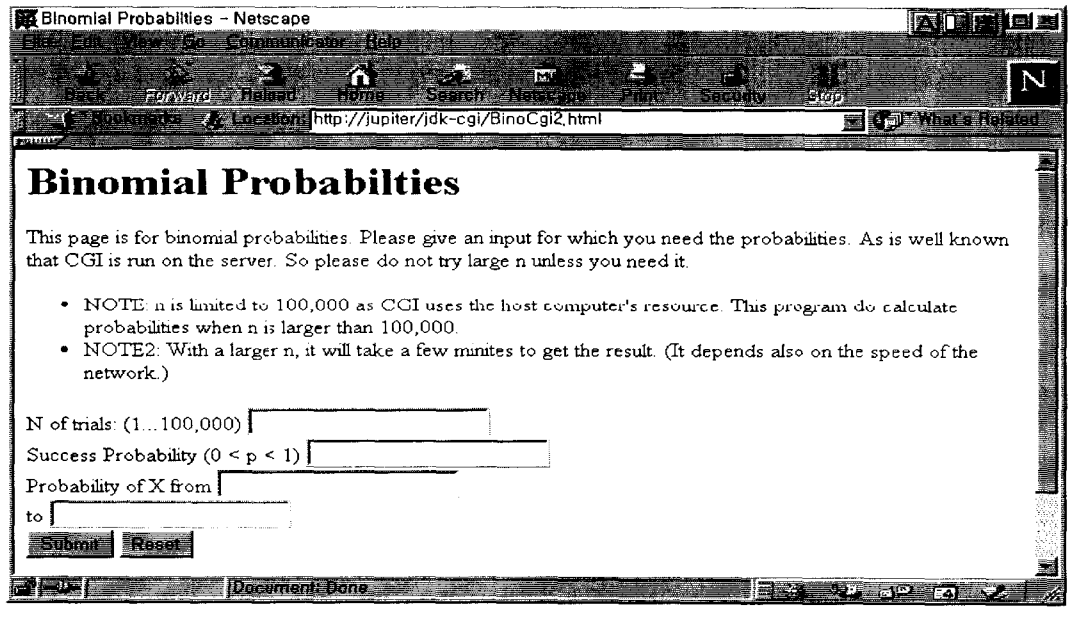
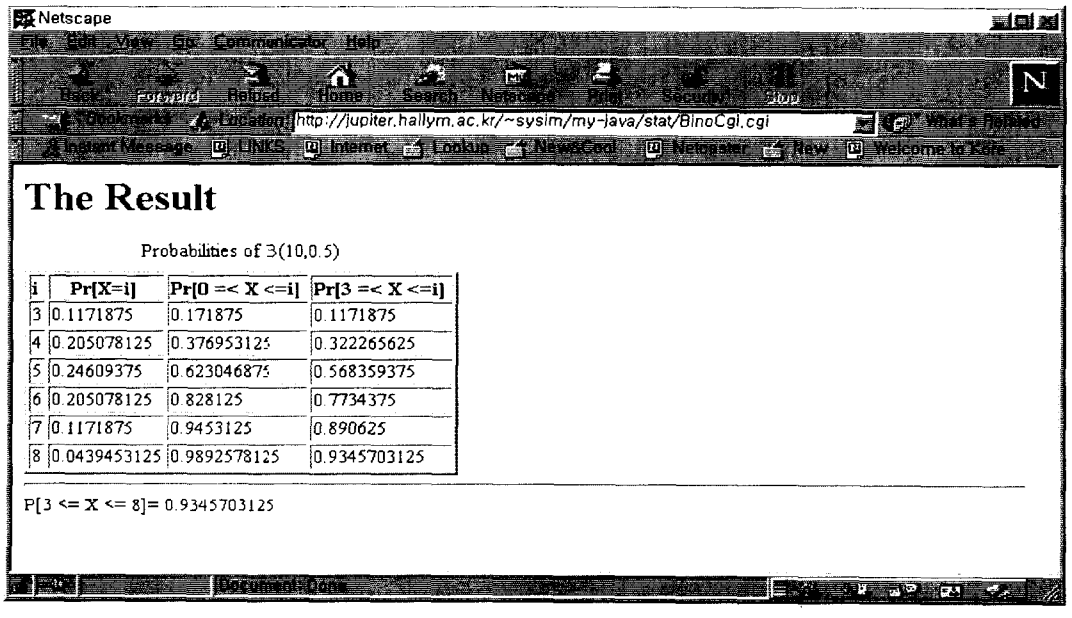


그림 4 : 이항 확률 CGI



i	Pr[X=i]	Pr[0 =< X <=i]	Pr[3 =< X <=i]
3	0.1171875	0.171875	0.1171875
4	0.205078125	0.376953125	0.322265625
5	0.24609375	0.623046875	0.568359375
6	0.205078125	0.828125	0.7734375
7	0.1171875	0.9453125	0.890625
8	0.0439453125	0.9892578125	0.9345703125

그림 5 : 이항 확률 CGI의 결과

전체 확률합이다. 이 CGI를 얻을 수 있는 장소는

http://jupiter.hallym.ac.kr/jdk-cgi/BinoCgi2.html

인데 초기 화면은 그림 4와 같다. 이 그림에서 보는 폼(FORM)에 필요한 값을 입력하고 Submit 버튼을 클릭하면 주어진 모수와 X의 범위에 해당하는 이항확률을 얻을 수 있다. 예를 들어 $n = 10$, $p = 0.5$, from에는 a 값으로 3, to에는 b값으로 8을 입력한 계산결과로 확률값 ($\Pr[X=x]$)과 그 누적합 ($\Pr[X \le x]$) 및 $\Pr[a \le X \le x]$ 값을 브라우저를 통해 사용자에게 전달하도록 하였다. 그림 5에는 이 결과를 담은 것이다. 포아송 확률의 계산은

http://jupiter.hallym.ac.kr/jdk-cgi/PoiCgi.html

에서 얻을 수 있는데 사용법은 앞의 이항분포의 경우와 마찬가지이다. 포아송 분포의 경우 확률변수가 취할 수 있는 값이 무한이므로 최대 0부터 3λ사이의 값에 대해서만 확률을 계산하도록 하였고 사용자가 입력한 시작과 끝의 사이의 값에 대해서만 확률과 누적확률 등을 계산하도록 하였다. 결과는 앞의 이항분포와 같은 모양의 출력인데 각각의 값에 대한 확률과 누적 확률 값을 얻을 수 있다.

6. 결론

본 연구에서는 포아송 확률과 이항 분포 및 정규분포 간의 근사에 대해 수치 및 동적 그림을 이용하여 분석하였다. 본 연구에서는 대부분의 컴퓨터 패키지에서 제공하지 못하는 값을 CGI를 통하여 얻을 수 있게 했으며 또한 이를 애플릿으로 만들어서 인터넷을 통하여 접근할 수 있게 하였다. 이러한 시도를 정리하면 살아 있는 전자교재가 될 수도 있을 것이며 통계를 사용하여야 하면서도 어려움 때문에 접근하지 못하는 많은 사람에게 유익한 참고가 되리라 생각한다.

참고문헌

[1] 강희모, 심송용 (1998), 웹상에서의 신뢰구간의 구현, "한국통계학회 경기강원인천 지회 학술발표회 논문집," 55-60. [2]

[2] 심송용 (1997), 인터넷을 이용한 원격수업, "1997년 한국통계학회 추계 학술 발표회 논문집," pp. 1-5.

[3] 안기수, 허문열 (1998), 멀티미디어와 통계 소프트웨어를 활용한 회귀분석 학습시스템, "응용통계연구," 제 11권 2호, 389-401.

[4] 이우리, 최현집 (1998), 웹에서 운영되는 그래프 모형을 이용한 동적인 분석 시스템. "한국 통계학회 논문집", 제5권 3호, 755-766.

[5] 조신섭, 송문섭, 이윤모, 성병찬, 윤영주, 이현부 (1998), 기초통계교육을 위한 통계패키지의 비교 연구 및 엑셀을 이용한 한글 통계패키지의 구현, "1998년 한국통계학회 춘계 학술발표회 논문집," 75-79.

[6] 한경수, 안정용, 강윤비 (1998), 통계학 교육을 위한 전자 교재의 활용, "응용 통계연구" 제 11권 1호, 5-12.

[7] Johnson, N. L. and Kotz, S. (1969), *Distributions in Statistics*, Wiley, New York.

[8] Liberman, G.J. and Owen, D. B. (1961),*Tables of the Hypergeomet-ric Distribution*, Stanford University Press.

[9] General Electric Company (1962), 0*Tables of the Individual and Cumulative Terms of Poisson Distribution*, D. Van Nostrand Company, Inc. Princeton, NY.

[10] Yilmaz M. R. (1996), The Challenge of Teaching Statistics to Non-Specialists., *Journal of Statistics Education.* Vol. 4.No. 1.http://www.stat.ncsu.edu/info/jse/v4n1/yilmaz.html