응용통계연구 제 **12**권 2호, *1999*년, *pp. 671-681*

이항분포의 정규근사에 대한 고찰

장대홍[1)]

요 약

이항분포의 정규근사는 중심극한정리의 한 예로서 자주 언급되는데 정규근사를 하기 위한 시행횟수 n과 성공율 p에 대한 판정기준들이 다수 제시되고 있는 데, 본 논문은 이러한 판정기준들에 대하여 제약조건의 강도와 평균오차한계를 비교, 검토하였다.

1. 서론

이항분포의 정규근사 문제는 거의 모든 통계학 교재에서 언급하는 바 다음과 같이 서술할 수 있다.

1. 이항분포의 정규근사 : n이 크고 p의 값이 0 또는 1에 가깝지 않은 경우 이항분포 B(n,p)는 평균 np, 분산 npq인 정규분포 N(np,npq)에 가까워진다(단, $q = 1 - p$이다.).

제 6차 수학과 교육과정에 따른 현행 고등학교 수학 교과서 수학I 에서도 이항분포의 정규근사를 다루고 있는데, 18종 모두 다음과 같이 언급하고 있다.

2. 이항분포의 정규근사 : 확률변수 X가 이항분포 B(n,p)를 따르고, n이 충분히 클 때 X는 근사적으로 정규분포 N(np,npq)를 따른다.

이항분포의 정규근사에 대한 2의 서술에서 알 수 있는 사실은 n에 대한 언급만 있고, p에 대해서는 아무 언급이 없다는 것이다. n에 대한 언급에서도 'n이 충분히 크다'라고 했는데 충분히 크다면 어느 정도 커야 되는지에 대하여 구체적으로 명시하지 않고 있다. 이항분포에서 시행횟수 n이 크고 $p = \frac{1}{2}$ 인 경우의 근사공식은 De Moivre에 의하여 유도되었고, Laplace에 의하여 일반적인 경우로 확장되었다. 이항분포의 정규근사에 대한 1의 서술 중 'n이 크고 p의 값이 0 또는 1 에 가깝지 않다'는 표현이 구체적이지 못하기 때문에 많은 통계학 교재들이 정규근사 판정기준들을 제시하고 있다.

본 논문은 통계학 교재들의 조사를 통하여 저자가 정리한 10가지 판정기준들의 제약조건의 강도와 평균오차한계를 비교하고자 한다. 2절은 10가지 판정기준들을 제시·정리하였고, 3절에서는 이 10가지 판정기준들 각각에 대하여 제약조건의 강도와 평균오차한계를 비교하였다. 4절에서는 결론을 내렸다.

1) (608-737) 부산광역시 남구 대연3동 599-1, 부경대학교 자연과학대학 수리과학부, 교수

2. 정규근사 판정기준들

통계학 책에 나타나는 10가지 정규근사 판정기준들을 정리하면 표 2.1과 같다. 이 표에서 알 수 있는 사실은 8번 판정기준($np \geq 5$, $n(1-p) \geq 5$)이 다른 판정기준들보다 월등하게 많이 언급되어 있고, 그 다음이 3번 판정기준 ($np(1-p) \geq 10$)인 것을 알 수 있다. 흥미로운 사실은, 판정기준이 10가지씩이나 제시되고 있지만, 문헌들에 나타나는 대부분의 예제들은 1번이나 2번 판정기준에 맞는 예들이라는 사실이다.

표 2.1: 10가지 판정 기준들

판정기준	제약조건	언급한 문헌
1	$np(1-p) \geq 20$	Roussas(1973)
2	$np > 15$ $n(1-p) > 15$	강 석복 외 3인(1997), 구 자홍 외 6인(1997), 김 동희 외 6인(1997)
3	$np(1-p) \geq 10$	Rohatgi(1984), Fisher와 Van Belle(1993), Ross(1997), Ross(1998)
4	$np(1-p) > 9$	Bowker와 Lieberman(1972), Aczel(1993)
5	$np > 10$ $n(1-p) > 10$	Kotz와 Johnson(1982)
6	$n > 9\max(\frac{1-p}{p}, \frac{p}{1-p})$	Larson(1995)
7	$np(1-p) \geq 5$	Rosner(1990)
8	$np \geq 5$ $n(1-p) \geq 5$	Gilbert(1976), Hoel(1976), Olkin, Gleser와 Derman(1980), Chilton(1982), Trivedi(1982), Bourke, Daly와 Mcgilvray(1985), Goldman과 Weisbeg(1985), Kotz와 Johnson(1985), 임 양택(1986), Bain과 Engelhardt(1987), 김 연형 (1988), 장 욱배와 김 윤기(1988), Godfrey, Roebuck와 Sherlock(1988), 김 상익 외 4인(1989), Glantz(1989), Lapin(1990), 이 외숙 외 3인(1991), Freund(1992), Milton(1992), Aczel(1993), Ott(1993), Pagano와 Gauvreau(1993), Walpole와 Myers(1993), Anderson, Sweeney와 Williams(1994), Creighton(1994), Jarrell(1994), Kelly(1994), Mason, Lind와 Marchal(1994), Krishnamurty, Kasovia-Schmitt와 Ostroff(1995), Mann(1995), Triola(1995), Weiss(1995), 강 근석 외 4인(1996), 김 우철 외 7인(1998)
9	$np \geq 4$ $n(1-p) \geq 4$	Mendenhall과 Sincich(1995)
10	$p \pm 2\sqrt{\frac{p(1-p)}{n}}$ $\in (0,1)$	Mendenhall, Sheaffer 와 Wackerly(1987), Sheaffer와 McClave(1990),Mendenhall과 Sincich(1992)

3. 판정기준들의 비교

이장택(1998)은 5가지 판정기준들(3, 4, 6, 8, 10번)에 대하여 시행횟수는 $10 \leq n \leq 50$, $n = 10, 11, \cdots, 50$으로 놓고, 성공율은 $0.05 \leq p \leq 0.95$로서 0.05 간격으로 놓고 시뮬레이

션을 행하였다. 연속성 수정을 고려하지 않는 경우와 고려한 경우로 나누어 $P[a \leq X \leq b](a < b, a = 0, 1, \cdots, n-1; b = 1, 2, \cdots, n)$의 값을 이항분포를 이용한 참값과 정규근사를 이용한 근사값을 구하고 이 차의 절대값을 구하여 5개의 판정기준들을 비교하였다. 이 방법의 단점은 판정을 위하여 너무 많은 경우 수를 고려하여야 한다는 것이다. 예로 제일 많은 경우 10번 판정기준을 위하여 350,935번의 경우 수를 계산하였고, 제일 적은 경우 3번의 판정기준을 위하여 75,370번의 경우 수를 계산하였다. 5가지 판정기준들을 위하여 총 1,101,050번의 경우 수를 계산하였다. 본 논문에서는 이 보다 훨씬 간단한 방법으로 2절에서 제시한 10가지 판정기준들을 비교하고자 한다. $X_1, X_2, \cdots, X_n$ 각각이 독립이고, 성공율이 p인 베르누이 확률변수들이라 하고, $S_n = \sum_{i=1}^{n} X_i$이고, $E(X_i) = \mu, Var(X_i) = \sigma^2$ 이라 하자. $E|(x_i - \mu)/\sigma|^3 < \infty$ 라면

$$|P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x) - \Phi(x)| \leq 0.7975E|X_i - \mu|^3/\sigma^3 \tag{3.1}$$

이 되고, 오차한계(error bound) 즉, 부등식 (3.1)의 오른쪽 값이

$$e(n,p) = 0.7975[1 - 2p(1-p)]/\sqrt{np(1-p)} \tag{3.2}$$

가 된다(Rohatgi(1984)). n이 자연수이나, 연속적인 값이라 놓아도 판정기준 제약조건의 강도 및 평균오차한계 분석에 영향을 미치지 않으므로 앞으로의 서술에서 n은 연속적인 값으로 여기기로 한다. e(n,p)의 모양을 수학 패키지 Maple로 그려보면 그림 3.1과 같다.

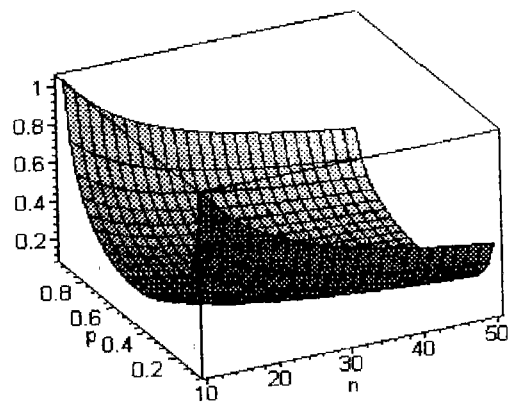

그림 3.1: e(n,p)의 모양

p=0.5를 중심으로 좌우대칭 그림이고, n이 커질수록 e(n,p)의 값이 작아지고, p에 따른

변화가 약하여진다. 10가지 판정기준들을 만족하는 (n,p)의 범위의 경계선들을 그려보면 그림 3.2와 같다(단, $1 \leq n \leq 100$).

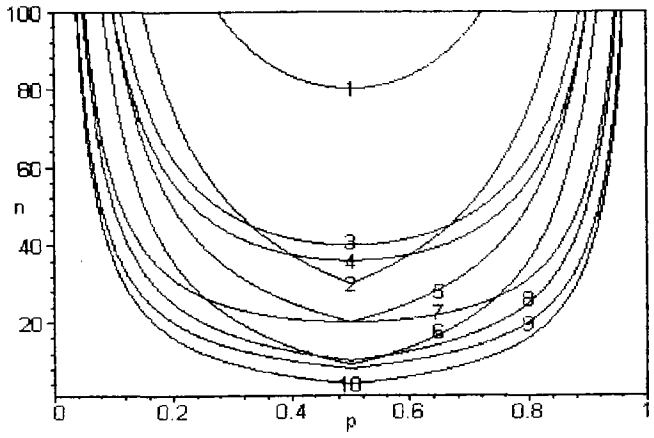

그림 3.2: 10가지 판정기준들을 만족하는 (n,p)의 범위의 경계선

이 그림을 통하여 10가지 판정기준들의 제약 조건의 강도를 알 수 있는데, 제약조건이 강한 순서대로 나열하면 1번 > 2번> 3번 > 4번 > 5번 > 6번 > 7번 > 8번 > 9번 > 10번 순이다. 여기서는, 제약조건을 만족하는 (n,p)의 범위의 크기를 가지고 제약조건의 강도로 삼았는데, 이 범위의 크기가 표 3.1에 나타나 있다. 이 범위의 크기는 이중적분을 이용하여 쉽게 구할 수 있는데 Maple을 이용하여 구하였다. 예로 1번 판정기준에 대한 범위의 크기는

$$\int_{1/2-\sqrt{5}/10}^{1/2+\sqrt{5}/10} \int_{\frac{20}{p(1-p)}}^{100} 1 dndp = 6.224$$

이다.

이 범위의 크기가 작을수록 제약조건의 강도가 크다. 제약조건을 만족하는 범위에서의 오차한계들의 평균값을 평균오차한계라 정의하고, 10가지 판정기준들의 평균오차한계를 Maple을 이용하여 구하니 표 3.1과 같았다. 이 값도 이중적분을 이용하여 쉽게 구할 수 있다. 예로 1번 판정기준에 대한 평균오차한계는

$$\frac{\int_{1/2-\sqrt{5}/10}^{1/2+\sqrt{5}/10} \int_{\frac{20}{p(1-p)}}^{100} 0.7975 \cdot \frac{1-2p(1-p)}{\sqrt{np(1-p)}} dndp}{\int_{1/2-\sqrt{5}/10}^{1/2+\sqrt{5}/10} \int_{\frac{20}{p(1-p)}}^{100} 1 dndp} = 0.089$$

평균오차한계가 작은 순서대로 나열하면 1번 > 2번> 3번 > 4번 > 5번 > 6번 > 7번 > 8번 > 9번 > 10번 순이다. 제약조건의 강도가 높을 수록 평균오차한계가 작았다. 즉, 제약조건

의 강도가 높을 수록 이항분포를 이용한 확률값과 표준정규분포를 이용한 확률값의 차이가 적게 된다.

표 3.1: 10가지 판정기준들의 범위의 크기와 평균오차한계

판정기준	범위의 크기	평균오차한계
1	6.224	0.089
2	33.881	0.114
3	36.191	0.119
4	40.450	0.123
5	47.811	0.129
6	58.582	0.143
7	60.570	0.150
8	66.974	0.158
9	71.794	0.168
10	75.480	0.175

이항분포의 정규근사에서는 통상적으로 근사계산의 정확도를 높이기 위하여 연속성 수정을 행하는데, 이 연속성 수정의 효과를 보기 위하여 $g(z) = P[z \le Z \le z + \frac{0.5}{\sqrt{np(1-p)}}]$ 값을 n=15, 100, 300 , p=0.05, 0.5에 대하여 $-5 \le z \le 5$ 범위에서 그려보니 그림 3.3과 같았다.

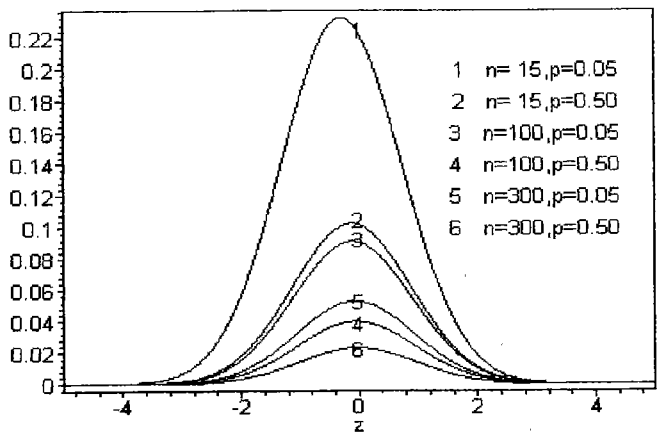

그림 3.3: g(z)의 값

이 그림에서 알 수 있는 사실은 p가 0이나 1에 가까울수록 연속성수정의 효과가 크고, p가 0.5에 가까울수록 연속성수정의 효과는 떨어진다. 또한, n이 커지면서 연속성수정의 효

과는 줄어들고, z의 위치가 0에 가까울수록, 즉 연속성수정을 행하는 위치가 0에 가까울수록 연속성수정의 효과가 크고, z의 위치가 0에서 멀수록 즉, 연속성수정을 행하는 위치가 0에서 멀수록 연속성수정의 효과는 급격하게 떨어진다.

4. 결론

이항분포의 정규근사에 대하여 10가지 판정기준의 제약조건의 강도와 평균오차한계에 대하여 조사하여 본 결과, $np(1-p) \geq c$(상수) 판정기준에서는 c가 큰 값일수록 강한 제약조건을 갖게 되고, ($np \geq c$(상수), $n(1-p) \geq c$)판정기준에서도 c가 큰 값일수록 강한 제약조건을 갖는다는 사실을 알 수 있었다. 제일 강한 판정기준은 $np(1-p) \geq 20$ 이었고, 제일 약한 판정기준은 $p \pm 2\sqrt{\frac{p(1-p)}{n}} \in (0,1)$이었다. 또한, 제약조건이 강할수록 이항분포를 이용한 확률값과 표준정규분포를 이용한 확률값의 차이가 적었다. 여기서 한가지 주의할 사항은 현재 컴퓨터의 발달로 표본의 크기가 크더라도 이항분포에서의 확률값은 정확히 구할 수 있다는 것이다. 그러므로, 이항분포의 정규근사가 주는 통계학적 의의는 그 근사값 계산에 있기보다 이산분포인 이항분포가 적절한 조건하에서 연속분포인 정규분포에 가까와진다는 사실 자체에 있다. 이산분포와 연속분포는 아주 다른 성질들을 갖고 있기 때문에 이산분포와 연속분포를 넘나들 수 있는 조건들을 따지는 것은 큰 의의를 지닌다.

참고문헌

[1] 〈고등학교 수학과 교육과정 해설〉(1992). 교육부.

[2] 강근석, 김성철, 김지현, 이윤오, 이정진, 이창수(1996). 〈PC통계학〉, 자유아카데미, 서울.

[3] 강석복, 오창혁, 우정수, 이광호(1997). 〈통계학〉, 형설출판사, 서울.

[4] 구자홍, 김진경, 박헌진, 이재준, 전홍석, 최지훈, 황진수(1997). 〈통계학〉, 자유아카데미, 서울.

[5] 금종해, 정순영, 박평순(1996). 〈고등학교 수학 I〉, 한샘출판, 서울.

[6] 김동희, 김충락, 손건태, 정광모, 정윤식, 최용석, 홍창곤(1997). 〈통계학〉, 자유아카데미, 서울.

[7] 김명령, 김창동, 박수화(1995). 〈고등학교 수학 I〉, 중앙교육진흥연구소, 서울.

[8] 김상익, 서한손, 안병진, 여성철, 이석구(1989). 〈통계학의 이해와 응용〉, 민영사, 서울.

[9] 김연식, 김홍기(1997). 〈고등학교 수학 I〉, 두산동아, 서울.

[10] 김연형(1988). 〈통계학 개론〉, 경문사, 서울.

[11] 김우철, 김재주, 박병욱, 박성현, 송문섭, 이상열, 전종우, 조신섭(1998). 〈통계학 개론〉, 영지문화사, 서울.

[12] 박규홍, 임성근(1995). 〈고등학교 수학 I〉, 동화사, 서울.

[13] 박두일, 신동선, 김기현, 박복현(1995). 〈고등학교 수학 I〉, 교학사, 서울.

[14] 박배훈, 정창현, 박상호, 류성림, 권기석, 류익승(1996). 〈고등학교 수학 I〉, 교학사, 서울.

[15] 박세희, 정광식, 강병개(1995). 〈고등학교 수학 I〉, 동아서적, 서울.

[16] 박한식, 구광조, 정지호, 이동수, 이강섭, 황선욱(1995). 〈고등학교 수학 I〉, 지학사, 서울.

[17] 양승갑, 이성길, 배종숙(1995). 〈고등학교 수학 I〉, 금성교과서, 서울.

[18] 우정호(1995). 〈고등학교 수학 I〉, 지학사, 서울.

[19] 윤옥경, 윤재한, 허원, 송병희(1997). 〈고등학교 수학 I〉, 중앙교육진흥연구소, 서울.

[20] 이병권, 박성규(1997). 〈고등학교 수학 I〉, 성안당, 서울.

[21] 이외숙, 임용빈, 성내경, 소병수(1991). 〈통계학〉, 경문사, 서울.

[22] 이장택(1998). 이항분포의 정규근사, 〈한국수학교육학회지 시리즈 A〉 , 제37권 제2호, 227-231.

[23] 이현구, 지동표, 김우철, 고성은, 박병욱, 장훈, 최용준(1997). 〈고등학교 수학 I〉, 천재교육, 서울.

[24] 이홍천, 강옥기, 박재석(1995). 〈고등학교 수학 I〉, 동아출판사, 서울.

[25] 임양택(1986). 〈통계학〉, 대명사, 서울.

[26] 장욱배, 김윤기(1988). 〈교양통계학〉, 일신사, 서울.

[27] 정봉화, 이우영, 신항균(1997). 〈고등학교 수학 I〉, 형설출판사, 서울.

[28] 정태환, 서태영, 유복동, 김광환, 박재명(1996). 〈고등학교 수학 I〉, 두산동아, 서울.

[29] 조승제(1995). 〈고등학교 수학 I〉, 재능교육, 서울.

[30] 조태근, 채윤기, 손규현, 김철언, 임성모, 정상권, 이재학(1997). 〈고등학교 수학 I〉, 금성교과서, 서울.

[31] Aczel, A. D.(1993). *Complete Business Statistics(2nd ed.)*, Irwin, Homewood.

[32] Anderson, D. R., Sweeney, D. J., and Williams, T. A.(1994). *Introduction to Statistics-Concepts and Applications(3rd ed.)*, West Publishing, St. Paul.

[33] Bain, L. J. and Engelhardt, M.(1987). *Introduction to Probability and Mathematical Statistics*, Duxbury Press, Boston.

[34] Bourke, G. J., Daly, L. E., and Mcgilvray, J.(1985). *Interpretation and Uses of Medical Statistics*, Blackwell Scientific Publications, London.

[35] Bowker, A. N. and Lieberman, G. J.(1972). *Engineering Statistics*, Prentice-Hall, Englewood Cliffs.

[36] Chilton, N. W.(1982). *Design and Analysis in Dental and Oral Research*, Praeger, New York.

[37] Creighton, J. H. C.(1994). *A First Course in Probability Models and Statistical Inference*, Springer-Verlag, New York.

[38] Fisher, L. D. and Van Belle, G.(1993). *Biostatistics*, John Wiley, New York.

[39] Freund, J. E.(1992). *Mathematical Statistics(5th ed.)*, Prentice-Hall, Englewood Cliffs.

[40] Gilbert, N.(1976). *Statistics*, W. B. Sounders, Philadelphia.

[41] Glantz, S. A.(1989). *Primer of Biostatistics* , McGrawHill, New York.

[42] Godfrey, M. G., Roebuck, E. M., and Sherlock, A. J.(1988). *Concise Statistics*, Edward Arnold, London.

[43] Goldman, R. N. and Weisberg, J. S.(1985). *Statistics: An Introduction*, Prentice-Hall, Englewood Cliffs.

[44] Hoel, P. G.(1976). *Elementary Statistics*, John Wiley, New York.

[45] Jarrell, S. B.(1994). *Basic Statistics*, Wm. C. Brown Publishers.

[46] Kelly, D. G.(1994). *Introduction to Probability*, Macmillan. New York.

[47] Kotz, S. and Johnson, N. L., ed.(1982). *Encyclopedia of Statistical Sciences Vol.1*, John-Wiley, New York.

[48] Kotz, S. and Johnson, N. L., ed.(1985). *Encyclopedia of Statistical Sciences Vol.6*, John-Wiley, New York.

[49] Krishnamurty, G. B., Kasovia-Schmitt, P., and Ostroff, D. J.(1995). *Statistics*, Bartlett Publishers, Boston.

[50] Lapin, L. L.(1990). *Probability and Statistics for Modern Engineering (2nd ed.)*, PWS-Kent, Boston.

[51] Larson, H. J.(1995). *Introduction to Probability*, Addison-Wesley, Reading.

[52] Mann, P. S.(1995). *Introductory Statistics*, John Wiley, New York.

[53] Mason, R. D., Lind, D. A., and Marchal, W. G. (1994). *Statistics: An Introduction(4th ed.)*, Harcourt, Brace & Company, Fort Worth.

[54] Mendenhall, W., Sheaffer, R. L., and Wackerly, D. D.(1987). *Mathematical Statistics with Application*, Duxbury Press, Boston.

[55] Mendenhall, W. and Sincich, T.(1992). *Statistics for Engineering and the Sciences(3rd ed.)*, Macmillan, New York.

[56] Mendenhall, W. and Sincich. T.(1995). *Statistics for Engineering and the Sciences*, Prentice-Hall, Englewood Cliffs.

[57] Milton, J. S.(1992). *Statistical Methods in the Biological and Health Sciences*, McGraw-Hill, New York.

[58] Olkin, I., Gleser L. J., and Derman, C.(1980). *Probability Models and Applications*, Macmillan, New York.

[59] Ott, R. L.(1993). *Introduction to Statistical Methods and Data Analysis*, Duxbury Press, Belmont.

[60] Pagano, M. and Gauvreau, K.(1993). *Principle of Biostatistics*, Duxbury Press, Belmont.

[61] Rohatgi, V. K.(1984). *Statistical Inference*, John Wiley, New York.

[62] Roussas, G. G.(1973). *A First Course in Mathematical Statistics*, Addison-Wesley, Reading.

[63] Rosner, B.(1990). *Fundamentals of Biostatistics*, PWS-Kent, Boston.

[64] Ross, S. M.(1997). *Introduction to Probability Models*, Academic press, London.

[65] Ross, S.(1998). *A First Course in Probability(5th ed.)*, Prentice-Hall, Englewood Cliffs.

[66] Sheaffer, R. L. and McClave, J. T.(1990). *Probability and Statistics for Engineers(3rd ed.)*, Duxbury Press, Belmont.

[67] Triola, M. F.(1995). *Elementary Statistics(5th ed.)*, Addison-Wesley, Reading.

[68] Trivedi, K. S.(1982). *Probability Statistics with Reliability, Queuing and Computer Science Applications*, Prentice-Hall, Englewood Cliffs.

[69] Walpole, R. E. and Myers, R. H.(1993). *Probability and Statistics for Engineers and Scientists(5th ed.)*, Macmillan, New York.

[70] Weiss, N. A.(1995). *Introductory Statistics(4th ed.)*, Addison-Wesley, Reading.

[1999년 1월 접수, 1999년 6월 최종수정]

A Study on Normal Approximation to the Binomial Distribution

Dae-Heung Jang[1)]

ABSTRACT

The central limit theorem enables us to calculate probabilities for a binomial random variable by approximating the binomial distributions with a normal curve. There are ten criteria for normal approximation to the binomial distribution. We compare these criteria with respect to degree of restriction and the mean of the error bound.

1) Division of Mathematical Science, Pukyong National University, 608-737, Pusan, Korea.