Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

통계계산 분야의 현황과 장래

김철웅[1], 심송용[2], 한경수[3]

1. 통계계산연구회 출현

1981년에 발간된 한국통계학회 창립 10주년 통계학연구 기념호의 통계와 전산분야에서 허문열 교수와, 토론자 박성현 교수가 통계계산에 관한 전문소위원회의 필요성을 제기하였다. 그리고 1988년 3월 31일 통계계산연구회의 창립총회를 갖게 되었으며, 현재까지 주기적으로 논문발표회 및 워크샵(Workshop)을 개최하며 통계계산 분야의 확산에 대한 노력을 경주하고 있다.

통계계산 분야의 필요성과 컴퓨터 및 정보기술(Information Technology)의 발전에 관한 일반적인 내용은 통계학회 10주년, 20주년 기념호 등에 잘 언급이 되어있다(허문열, 1981; 김병천, 1991).

2. 정보기술의 발달과 시대적 변화에 따른 통계계산 분야의 변화 양상

"Statistical computing touches on almost every aspect of statistical theory and practice, and at the same time nearly every aspect of computer science."
라는 Thisted(1988)의 말처럼 통계계산은 전산학 분야와 유기적으로 연결되어 있어 전산학 분야의 발전과 그 행보를 같이 하고있다. 전산분야의 발전은 또한 컴퓨터의 발전과 그 행보를 같이 할 것이고 컴퓨터의 발전은 곧바로 통계계산 분야에 지대한 영향을 주고 있고, 또한 앞으로 줄 것이다.

전산학 또는 컴퓨터의 발전은 새로운 통계방법을 탄생시켰다. 80년대 전후의 비모수분야(Bootstrapping, Nonparametric density estimation 등)의 발전이 그러하였고, 90년대의 통계그래픽 분야(dynamic statistical graphics, representation of high dimensional data, Tomography 등)의 괄목할만한 발전은 컴퓨터의 발전에 기초한 것이다. 90년대 후반의 베이지안 들의 계산에 대한 고민을 해결해 준 것(Monte Carlo Markov Chain등)도 모두 컴퓨터의 발전에 기초하였다.

자료분석적인 측면에서 본다면 초기 단순반복작업이 가능한 컴퓨터 시대는 성능과 비용 등의 제약으로 데이터의 단순한 집계와 요약을 위주로 통계가 사용되었다. 1980년대 말에는 컴퓨터의 고성능화와 통계분석을 위한 PC용 패키지(SAS, SPSS 등)들이 속속 등장하게 되었다. 또한 개인용 컴퓨터의 보편화는 많은 통계패키지의 보급을 촉진시켜, 여러 분야에서 통계자료 분석의 필요성과 통계상담의 요구를 크게 증가시켰다. 1990년대 중반에 등장한 웹(Web)은 그 편리성과 유용성으로 국내 통계학 연구의 전반에 영향을 주었는데 특히 통계계산, 통계교육, 조사통계 등에서 웹을 이용하는 연구가 활발히 진행되어 왔다. 1990년대 말에는 고성능의 개인용 컴

1) 연세대학교 응용통계학과
2) 한림대학교 통계학과
3) 전북대학교 통계학과

퓨터와 편리하고 고급화된 정보기술(프로그래밍 언어, 운영체제의 발달, 데이터베이스의 보편화 등)이 보편화되었다. 이로 인해 사회 일반의 여러 분야에서 많은 양의 데이터가 발생·축적되어 졌고, 대용량의 데이터를 분석하기 위한 통계계산 분야의 필요성이 증대되었다. 특히 한국산 데이터 마이닝 도구의 개발, 통계 데이터베이스의 활용, 한국형 통계패키지의 개발, 그리고 CRM 등과 관련된 연구는 사회 일반에 통계학, 특히 통계계산 분야에 대한 연구와 전문인력 양성의 중요성과 필요성을 인식시키는 계기가 되었다고 할 수 있다.

3. 통계계산의 현황

지난 10년간은 컴퓨터 하드웨어의 발전과 인터넷의 보편화로 인해 컴퓨터에 대한 일반의 인식조차 바뀌었다. 이러한 발전은 통계학의 각 분야에도 다양한 영향을 주었다. 컴퓨터의 연산속도가 빨라지면서 이전에는 시간과 비용 때문에 실행하지 못했던 분야들이 보편화되기도 하였다. 이를 몇가지 분야로 나누어 보면 다음과 같다.

Data Augmentation, Markov chain Monte Carlo, Gibbs Sampler

수치적분조차 적용하기 어려워서 이전까지는 적분형태로 주어지던 posterior 분포의 기대값이나 HPD(Highest Posteror Density)등을 조건부 분포의 난수를 아주 긴 Markov chain 형태로 생성하여 그에 대한 emperical 기대값나 HPD를 얻을 수 있게 되었다. 이러한 방법은 하나의 난수를 얻기 위해서도 긴 연쇄를 이용하기 때문에 많은 난수의 생성을 요구하게 되고 따라서 연산속도가 뒷받침되지 않는 경우에는 현실적으로 사용할 수 없었다. 이들 방법의 기초가 되는 Metropolis 알고리즘(Metropolis(1953))이나 EM 알고리즘(Demster et.al.(1977))이 통계학 분야 소개된 한참이 지난 후에 재조명할 수 있게 된 것은 컴퓨터의 발달에 기인한 것이었다.

이러한 일련의 방법들은 Tanner(1993), Geman and Geman(1984), Gelfand and Smith(1990) 등의 연구에 의해 발전되었고 이러한 연구는 Bayesian 뿐만 아니라 frequentist들에게도 유용한 방법으로 도입되어 유전자 정보 분석이나 다음에서 보는 신호처리 등에서 직접적으로 응용되고 있다.

통계적 신호처리

CD, 디지털 카메라, 디지털 캠코더 등은 이제 일반 가정에서도 흔히 접할 수 있는 가전제품에 속한다. 이러한 매체들은 대부분 큰 용량의 메모리를 필요로 하고 그렇기 때문에 대부분 압축기술을 사용하게된다. 따라서 위성사진, 음성신호 등의 잡음(noise)을 처리하는 등의 현실적 필요에서 시작한 신호처리 분야는 Besag(1986)의 예에서 보는 잡음의 제거 뿐 아니라 동영상에서 움직이지 않는 부분에 대한 처리를 하지 않기 위해 사용하는 edge detection 기술, CT와 MRI로 대표되던 영상의료기기의 발전된 기술인 PET(positron emission tomography) 등의 기술적 분야에 통계학이 응용되고 있다.

다양한 통계 패키지

잘 알려진 통계 패키지인 SAS나 SPSS와 같은 범용 패키지는 많이 사용하고 있고 대부분의 통계적 방법을 제공하기 때문에 특수한 경우를 제외하고는 솔루션을 제공해 왔다. 그러나 CPU의 발전에 따라 정확한 유의확률을 구하기 어려웠던 분야에도 permutation test 등을 이용하여 정확한 유의확률의 계산을 짧은 시간에 할 수 있게 되었고 이러한 패키지들도 얻을 수 있게 되었다. 대표적인 예로 StatXact

http://www.statcon.de/vertrieb/produkte/STATX.HTM

가 있다.

Data base와 Data mining

현대의 갈수록 다양하고 매시간 쏟아지는 새로운 자료들이 데이터베이스에 저장되고 있으나 이들 자료들에서 구체적이며 유의미한 패턴을 끌어내는 데는 어려움이 많았으나 데이터 마이닝이 소개되고 컴퓨터의 연산속도가 빨라짐으로서 기가바이트 단위의 자료를 처리할 수 있게 되었다. 이러한 상황에서는 통계학자에게도 데이터 베이스 등의 컴퓨터에 관련 더 많은 지식을 요구하게 되었다.

이러한 현실에서 1998년 이후 금융기법의 선진화 노력이후 요구되는 위험관리, Credit scoring, CRM(Customer Relationship Management), 보험이나 신용카드의 Fraud detection 등 매우 다양한 분야에서 컴퓨터와 통계학에 대한 양쪽 지식을 요구하고 있다.

Web과 통계

1990년대 중반이후부터, 구체적으로는 웹브라우저의 보편화로 인하여 이전에는 특수한 계층의 사람들에게만 사용되던 각종 인터넷이 일반인들에게까지 널리 보급되었고 이제는 학술연구에서도 인터넷이 없어서는 안될 존재가 되었다. 이러한 현상은 웹서비스를 이용한 각종 통계의 제공에서부터 통계학 개념의 설명에까지 이용되고 있으며 이러한 서비스는 국내외에서 얻을 수 있다. 이 서비스는 많은 경우에 자바 애플릿을 이용하여 제공되는데 예를 들면 http://anova.hallym.ac.kr 등에서 보기를 볼 수 있다. 인터넷에서 통계관련 계산을 해주는 사이트의 목록은

http://members.aol.com/johnp71/javastat.html

등에서 얻을 수 있다. 자바 애플릿은 서버의 CPU를 사용하지 않고 클라이언트의 CPU를 사용하기 때문에 많은 서비스 요청에도 서버에 과부하가 걸리지 않은 장점이 있기 때문에 웹에서도 각광받는 언어이다.

인터넷 여론조사

인터넷의 대중화에 힘입어 각종 언론 매체 등에서 특별한 비용을 들이지 않고 조사를 하기 위한 목적으로 웹사이트에 여론조사 항목을 넣기 시작하였다. 이러한 방식의 조사는 확률표본이 아니기 때문에 오차한계나 신뢰도 등의 관점에서 보면 많은 문제가 있다. 따라서 인터넷 여론조사에 관련된 분야에 대한 연구가 필요한 시점이다.

통신과 통계계산

이젠 전화가 음성만 전달하던 시대가 아니다. 특히 mobile phone의 경우 모두 디지털로 데이터가 전송되며 IMT2000 이후 동영상의 전달이 필요하게 됨에 따라 동영상과 음성의 동시 전달로 인해서 제한된 용량 내에서 소비자가 만족할 수 있는 음성과 영상의 기준점을 찾아야할 필요가 있다. 즉, 음성 부분의 품질을 좋게 하면 동영상의 화질이나 움직임이 크게 떨어지고 동영상의 품질을 올리게 되면 음성 부분의 품질이 떨어지는 trade-off가 생기는데 이를 위한 관능검사 등의 통계적 기술이 필요하게 된다.

통신 트래픽은 엄청난 양의 자료를 생산해 내고 있는데 최대통신 트래픽은 교환기의 증설 필요성에 대한 중요한 자료가 된다.

4. 누적되어온 문제점

학회 10주년, 20주년 기념호에서 강조하였던, 사회 일반의 요구를 반영하는 응용통계 분야에

대한 노력이 여전히 부족한 상황이다. 통계계산 분야의 교육도 마찬가지로 단순한 프로그램 언어의 교육 등에 그치거나 외국산 통계패키지의 활용방법의 교육 정도에 그치고 있어, 실제 현장에서 필요로 하는 응용통계, 특히 통계계산 분야에 대한 인력의 양성과 배출은 턱없이 부족하다고 하겠다. 또한 급격히 발전하고 있는 정보기술(IT)을 예견하고 이에 대비하는 능력이 부족하여 이를 반영하는 통계계산 분야의 교과과정 개편이 잘 이루어지지 않고 있으며, 통계계산 분야의 전문강의 인력, 통계계산실습실, 실습예산 등의 부족이 통계계산 분야의 전문인력 양성에 걸림돌이 되고 있다.

5. 통계계산의 미래

컴퓨터의 계산 속도는 계속 빨라질 것이고, 기억 용량은 한없이 늘어만 갈 것이다. 현재는 불가능하다고 생각되어지는 것들이 빠른 계산 속도와 커다란 기억 용량을 가진 컴퓨터의 이용으로 통계 계산 분야의 새로운 영역은 계속 나타나고 늘어가게 될 것이다.

Wavelet이나 유전자 알고리즘과 같이 타 분야에서 계발된 새로운 개념이나 자료분석방법이 통계계산 분야에 영향을 미치고 통계학에 이전되어 기존의 방법과 융합되거나 다른 하나의 대안적 방법으로 자리를 잡아가는 현상은 계속될 것이다.

최근에 폭풍처럼 밀어닥친 계열별, 학부제 등의 학제 개편 열풍과 정보시대의 유행에 따라 '정보'라는 단어가 통계학과의 접두어가 되었다. 이것을 기회로 잘 이용하여 통계계산분야의 교과과정을 확립하여 학생들의 관심을 고조시키고, 취업한 후 통계계산 분야와 연관된 일을 할 수 있는 기회를 점점 넓혀간다면 고무적인 일이 될 것이다.

"As computer become more powerful, humans - for example statisticians - will, and must, continue to adapt by emphasizing in their training the things that computers cannot do."
라는 Porter(2001)의 말은 넓은 의미에서의 통계학자들의 역할에 대하여 언급한 것이지만, 이것을 통계계산에 국한하여 해석한다면 통계계산 분야의 통계 비전공자에게 유용성을 널리 알리고 통계계산의 방법을 통계나 계산을 모르는 사람들에게 쉽게 설명할 수 있도록 우리가 준비가 되어 있어야 하고 교육을 통하여 학생들도 그와 같은 준비가 되어있도록 가르쳐야 하겠다.

최근에 CRM(Consumer Relation Management), Financial Engineering, Risk Management, Data Mining 등 회사에서 실제 이러한 분야의 팀이 구성되어 있는 곳이 많아지고, 이러한 분야에 대한 consulting 업무를 하는 회사들도 많이 생겨나고 있는 추세이며, 이러한 추세는 계속될 것으로 전망된다. 사회 요구에 부응할 수 있는 교과과목을 개설하여 학생들이 통계계산분야에 대한 전문성을 갖고 취업을 할 수 있도록 하여야 할 것이다.

컴퓨터의 발전과 통계계산 분야에서의 영역확장은 그 맥을 같이한다. 컴퓨터의 발전이 끊임없이 이루어 질 것이라 전망하는 한, 통계계산 분야의 영역 또한 끊임없이 발전하게 될 것이다. 타 분야에서 개발된 방법을 통계적으로 해석하여 통계학의 한 분야로 만드는 작업을 게을리 하지 말아야 할 것이며, 통계계산의 고유한 분야에 대한 연구개발 또한 게을리 하지 말아야 통계계산의 미래가 있을 것이다. 미래는 우리가 계획한 방향으로만 가지 않고 의외의 모습으로 나타날 지라도 우리가 노력하는 만큼의 결실을 얻을 것이므로 통계계산 분야의 연구와 교육을 계속 경주하여야 할 것이다.

참고문헌

허문열(1981), "한국통계의 현황과 장래-통계와 전산", 10권, 창립 10주년 기념호, 77-80.
김병천(1991), "통계계산분야의 현재와 미래", 20권, 창립 20주년 기념호, 105-110.

Besag, J.E. (1986) On the Statical Analysis of Dirty Pictures, *JRSS* Ser. B, **48,** 259-302.

Demster, A.P., Laird, N. and Rubin, D.B. (1977) Maximun Likelihood from Incomplete Data via EM Algorithm(w discussion), *JRSS*, Ser. B, **39,** 1-38.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling Based Approaches to Calculation Marginal Densities, *JASA*, **85,** 398-409.

Geman, S. and Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions and the Baysian Restoration of Images, *IEEE transactions on Pattern Analysis and Machine Intelligence PAMI-6* 721-741.

Metropolis, N., Rosenbluth, A.W., Teller, A.H. and Teller, E. (1953) Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, **21**, 1087-1091.

Porter, T.M. (2001) "Statistical Futures", *Amstat News*, September 2001. pp.61-64.

Tanner, M.A. (1993) *Tools for Statistical Inferences,* Springer-Verlag, New York.

Thisted, R.A. (1988) *Elements of Statistical Computing*, Chapman and Hall.