통계계산분야의 현재와 미래

김 병 천*

1. 서론

통계계산, Statistical Computing, 또는 Numerical Computation for Statistics라고 불리우는 통계의 한 분야는 컴퓨터 산업에 의한 제 3의 산업혁명이 일고 있는 현대에서 매우 중요한 위치를 차지하고 있음은 두 말할 것도 없다. 특히 통계학이 컴퓨터가 발전하면서 더욱 더 학문적 발전에 박차를 가하고 있다는 것도 의심할 여지가 없다. 통계가 다루는 데이타는 적은 양으로부터 방대한 양을 다루고 있기 때문에 컴퓨터는 필수불가결한 파트너가 되었다. 1980년 초반에 16비트 컴퓨터가 개발되면서, 통계를 처리하는 장소가 전자계산소로부터 사무실로 옮겨 오기 시작했고, 최근에는 70MIPS이상의 속도를 갖고 있는 탁상용 Workstation이 개발되어 통계학자들의 마음을 설레게 하고 있다. 또한 대량의 데이타를 저장할 수 있는 Laser Compact Disk들이 개발되어 통계분야 및 통계계산분야의 발전에 기대가 더 모아지게 되었다. 그러면 컴퓨터가 발전되고 있는데 왜 통계계산분야의 발전이 필요하며, 현재까지 통계계산분야는 어느 단계까지 와 있으며, 미래를 위한 통계계산분야는 어떻게 변화할 것인가를 한국의 실정에 기초를 두고 논해 보고자 한다.

2. 통계계산분야의 필요성

통계를 다루는 많은 통계인들조차도 통계계산분야를 단순히 컴퓨터를 이용한 자료의 통계적 처리로 인식을 하고 있다. 그러나, 통계계산분야는 통계학과 전자 계산학을 보완적으로 이용하는 학문의 한 분야이다. 이 분야의 발생을 알아보면 컴퓨터는 계산상에 너무나도 헛점이 많아 우리가 원하는 계산의 결과를 얻을 수 없다는 것이 통계계산분야의 발생 원인이다. 간단한 예로서, 컴퓨터가 다룰 수 있는 수(number)에는 한계가 있음을 우리는 알고 있다. 즉, 모든 정수, 유리수, 실수를 다룰 수 없기 때문에 무한한 수를 다루고 있는 수학적 학문분야에는 적합하지 않다. 그 이유는 컴퓨터에서 사용

* 한국과학기술원 응용수학과

하고 있는 Register들의 길이가 한정되어 있기 때문이다. PC에서 수치보조장치(numeric data coprocessor)를 사용하면 그 길이가 80Bit이기 때문에 컴퓨터가 다룰 수 있는 실수의 개수는 그 조합을 생각해 보면 2^{80}개이다. 이 결과를 보면 아무리 우수한 슈퍼컴퓨터를 사용하더라도 이와 같은 한계성을 벗어날 수가 없음을 반드시 알아야 한다. 실제의 예를 들어 보면, 통계에서 많이 사용되는 표본분산식은 다음과 같다.

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

$$\bar{X} = (1/n) \sum_{i=1}^{n} X_i$$

여기서, 다음의 두 식은 수학적으로 똑같은 식임에 틀림없다.

$$(n-1)S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (2.1)$$

$$= \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \qquad (2.2)$$

그러나 컴퓨터 언어에서 Single Precision을 사용하여 각각의 결과를 구해 보면 그 결과가 틀린 경우를 볼 수 있다[1]. 그 이유는 무엇이며 (2.1)식과 (2.2)식중 어느 식이 정확한가? 해답은 (2.1)식이 정확한 식이다. 그러나, 전산학을 전공한 프로그래머의 입장에서 보면 (2.1)식은 데이타를 2번 읽어야 하는 단점이 발생한다. 이와 같은 문제를 해결해야 하는 학문이 바로 통계계산분야이다. SPSS와 SAS도 (2.1)과 (2.2)식중 어느 것도 사용하고 있지 않음을 알아야 한다. 사용하는 계산기법은

$$(n-1)S^2 = \sum_{j=2}^{n} \frac{(X_j - \sum_{i=1}^{j} X_i)^2}{j(j-1)}$$

이며, 계산 결과는 (2.1)과 같으며, 데이타를 한번 처리하여 얻을 수 있음을 알 수 있다. 다음의 예로서 표준 정규분포의 확률값을 구하고자 할 때, 즉

$$\int_0^{1.5} \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2} dx$$

의 계산방법도 쉽지 않음을 알 수 있다. 일반적으로 Quadrature방법을 많이 사용하고 있지만[2], 다변량 분석에서는 적분 자체가 많아지고, 복잡해지기 때문에 계산중 Error가 많이 발생되고 있어 그 결과를 얻기가 쉽지 않다. 이런 경우에도 통계계산기법을 이용하여 구하여야 한다. 이외에도 난수(random number)를 이용하여 확률변수(random variable)을 구하는 방법도

통계계산분야의 한 분야이다. 특히 이 분야의 새로운 기법 개발이 시뮬레이션을 발전시키는 한 분야임을 반드시 알아야 한다. 또한 회귀분석이나 실험계획법에서 많이 사용되는 역행렬의 계산방법도 역시 통계계산분야에서 개발해야 하는 한 분야이다. 많은 통계학 교과서나 논문에서 컴퓨터를 이용한 행렬의 계산부분에서도 틀린 부분이 지적되고는 한다.

3. 현재의 통계계산분야

대한민국에 컴퓨터가 처음 도입된 기관은 1966년 통계를 전문으로 다루는 통계청이었으며, 그 이후 많은 연구소들과, 대학교, 그리고 산업체에서 컴퓨터를 도입하여 사용하였지만, 통계계산분야, 수치해석분야, 소프트웨어 공학분야에 기여를 전혀하지 못하고 외국의 기법들을 도입하여 사용하는데만 급급해 왔다. 1980년부터, 통계계산분야의 필요함이 서구로부터 받아들여지기 시작했고, 현재는 이 분야에서 종사하는 통계인이 조금씩 늘어가고 있는 실정이다. 통계학의 기본원리를 가르치는 확률에서도 주사위를 던지는 실험을 하고 있지만, 국내의 어느 통계 교과서에도 컴퓨터를 이용한 주사위 던짐에 관한 시뮬레이션의 예제가 없는 실정이다. 또한 통계학의 기본법칙인 중심극한정리(central limit theorem)도 포본의 수가 많을 때 정말 정규분포를 하고 있는가의 실제적인 시뮬레이션에 관한 예제도 없다. 최근 미국에서 사용되고 있는 Calculus책들은 대학의 저학년부터 컴퓨터를 이용하도록 연습문제에 컴퓨터의 응용문제를 만들어 학생들에게 컴퓨터의 실습을 강조하고 있다. 통계학과의 교과 편람을 보더라도 통계계산분야의 과목이 설정되어 있지만 체계적으로 교육과정에 대하여서는 정립되지 않고 있음을 우리 통계인은 인식하고 있다. 대부분의 통계학과 졸업생들이 취업하는 분야도 한정되어 있고 대부분이 컴퓨터에 관한 분야에 종사를 하게 된다. 이들이 컴퓨터분야에 종사하면서 많은 어려움을 겪게 되는데, 제일 큰 문제는 컴퓨터의 프로그램 작성이 아닌, 통계계산분야의 문제점에 부딪치게 된다. 예를 들면, 통계패키지는 다양하게 다룰 수는 있지만, 통계패키지 안에 있는 통계분석과정의 알고리즘의 차이점, 분석하고자하는 통계적 방법이 없을 때 현재의 자료를 그대로 이용하여 정확한 결과를 얻을 수 있는 프로그램의 기법등이다. 이러한 문제점은 어디에서 오는 것일까? 첫째로 대학의 교과과정이 컴퓨터를 이용한 실습 위주보다는 강의 위주인 점이다. 저학년부터 강의를 겸한 실습교육이 필요하고 고학년에서는 실습 위주의 교과과정이 필요하다. 현재 몇몇 대학을 제외하고는 교과과정이 이론에 중점을 두고 있는 실정이다. 둘째는, 컴퓨터의 실습실부족을 들 수 있다. 많은 대학의 통계학과들이 예산 부족에 처해 있고, 대학 당국의 지원을 받지 못해, 학생들이 컴퓨터를 접해보는 시간이 매우 적은 편이다. 셋째로, 통계계산분야의 전문교수의 부족을 들 수 있다. 전문교수의 부족이 통계학을 전공하는 학생들의

컴퓨터 기피 현상을 초래하고 있다. 넷째로, 산업체와 기업에서의 인식 부족이다. 이 문제는 통계계산분야뿐만 아니라, 통계의 전 분야에 이르는 심각한 문제이다. 가까운 일본만 하더라도, 제약회사들은 통계계산분야를 전문적으로 활용하고 있으며, 각 회사마다 통계분석에 관한 실을 만들어 최소한 10명이상의 전문적 통계인 또는 통계계산인이 통계를 담당하도록 되어 있다. 특히 국내에서 통계의 활용범위가 큰 보험회사들도 통계에 대한 무관심은 커다란 문제로 대두된다. 다섯째로, 정부의 적극적 정책배려에 대한 무관심이다. 이 문제도 통계 전분야에 걸친 문제이지만 정확한 통계 처리를 위한 정부의 정책 설정이 안되고 있는 현실이다. 최근의 공해문제가 좋은 예가 될 것이다. 이상과 같은 국내의 현실에서 미래의 발전방향에 대하여 어떻게 대처해야 하는가?

4. 통계계산분야의 발전을 위한 제안

컴퓨터의 급속한 발전이 우리 통계인을 채찍질을 하고 있다. 현재의 모든 산업이 컴퓨터를 근거로 발전되고 있다는 것은 틀림없는 사실이다. 앞으로 10년 이내에 현재의 슈퍼 컴퓨터 성능을 갖고 있는 탁상용 컴퓨터가 통계인 각자의 앞에 놓여질 것이며, 모든 업무를 이 컴퓨터를 이용하여 처리를 하게 될것이다. 과연 우리 통계인들은 컴퓨터를 어느 정도 알고 있으며, 통계계산분야에 얼마나 관심을 갖고 있는가? 보다 더 중요한 것은 통계학 중 통계계산분야가 컴퓨터를 이용한 산업의 발전에 반드시 필요한 분야인 것을 알아야 한다. 앞으로 예상되는 분야는 다음과 같다.

가. 의학분야

의학분야중 임상 데이타처리와 분석에서의 통계학 요구는 다양하다. 그러나 통계적 자료처리보다는 인공지능 또는 전문가 시스템에서 통계계산분야의 적극적인 지원이 필요하며, 특히 X-ray필름을 대신하여 그래픽을 이용한 환부의 진단등에 필요하다.

나. 데이타베이스분야

데이타베이스분야는 통계계산분야의 도움없이 발전하기가 힘들다. 방대한 양의 데이타보다 사용자들은 축소된 데이타, 즉 통계량을 요구하기 때문에 통계분야와 통계계산분야가 필요하다.

다. 통신분야

근래 선진국을 비롯하여 급속히 발전되는 분야이며, 통신을 위한 프로토콜의 개발에 통계를 필요로 하고 있다. 통신하고자 하는 데이타를 암호화하여 보내고 이 데이타를 원래의 상태로 환원시키고자 할 때 통계계산분야의 연구가 필요하다.

라. 그래픽분야

컴퓨터에서 그림을 화면에 나타내거나, 레이져 프린터에 선을 긋고자 할 때 통계학과 통계계산분야가 필요하다. 특히 3차원 그래픽의 표현에서는 통계계산분야가 더욱 더 필요하다.

마. Signal Processing분야

레이다에 표시되는 자료의 분석등에 통계를 필요로 하는 최첨단 분야이다.

바. 자료 안전관리 분야

현대의 정보화하는 사회에서 컴퓨터의 이용은 오히려 부작용을 일으키는 경우가 많다. 특히 자료의 안전을 위한 비밀코드등의 개발에 통계계산분야의 도움을 필요로 하고 있다.

이상의 분야들은 통계계산분야의 도움없이 발전할 수가 없다. 이밖의 첨단 산업분야 이외에도 공학, 경제학, 심리학, 사회학 등에서도 통계와 통계계산분야를 필요로 하고 있다. 이와 같이 모든 분야에서 필요로 하고 있는 이 분야의 발전을 위해 다음과 같은 제안을 하고자 한다.

* 컴퓨터를 이용한 통계 실습에 의한 통계 교과목 개정

통계학 자체가 데이타를 이용하여 처리와 분석을 하고 있기 때문에 현재와 같이 컴퓨터가 발전하여 컴퓨터를 이용해야 하는 환경에서는 통계의 실습이 많이 필요하게 되었다. 졸업학년의 경우, 저학년에서 배운 통계의 이론을 컴퓨터를 사용하여 컴퓨터에 대한 충분한 실습을 반드시 해야 한다.

* 컴퓨터 실습에 관한 과감한 투자의 필요성

컴퓨터를 이용한 통계 교과목의 개정이 이루어지더라도, 컴퓨터의 하드웨어와 소프트웨어에 과감한 투자가 뒤따라야 한다. 컴퓨터만 구입되었다고 해서 컴퓨터를 사용할 수 있는 것이 아니고, 이에 따른 주변 부속장치와 응용 소프트웨어의 공급이 필요하다. 한 예로서, 만약 슈퍼컴퓨터가 도입되었다고 가정하자. 대부분의 사람들은 이 슈퍼컴퓨터 한대가 모든 문제를 해결한다고 믿는데, 슈퍼 컴퓨터는 단지 수치를 빠르게 처리하는 컴퓨터에 불과하지, 화일이나 Networking문제까지 해결할 수 없음을 알아야 한다.

* 부전공의 실시

통계를 전공하는 학생들은 반드시 부전공을 실시하여, 통계 및 통계계산분야를 필요로 하는 학문의 기본적인 원리를 알고 이해를 해야 한다. 특히 전산학의 기본적인 학문의 이해는 말할 필요조차 없다.

* 산학연으로 구성된 Working Group의 활성화

아무리 대학에서 통계계산분야가 활성화되더라도, 이 분야를 이용하여 사용하고 있는 산업체나 연구소에서 활성화가 안되면 무용지물이 된다. 통계인들 자신들이 기업과 산업체를 끌어들여 각종 세미나를 개최하고, Working Group을 활성화하여 보급하는데 힘써야 한다.

* 통계에 관한 정부의 과감한 정책 배려

통계는 정책을 결정하는 학문이 아닌, 정책의 결정을 도와주는 학문이다. 대부분의 사람들은 이러한 관계로 통계는 누구나 할 수 있다고 하지만, 통계학 자체가 어려운 학문임을 알아야 한다. 이러한 통계에 대한 가벼운 생각이 정부로부터 시작되기 때문에 통계를 일종의 데이타의 조작으로 많이 여기고 있다. 일본의 경우, 통계를 정부가 신뢰하고 있으며, 특히 제약회사의 경우 통계적 실험을 반드시 하도록 정부의 지침사항으로 넣어, 제품의 고급화를 꾀하고 있다. 정부는 통계를 널리 보급하고, 이에 따른 정책을 배려하여 온 국민이 통계를 생활화하도록 힘써야 한다.

현재와 같이 국내에서 통계가 외면을 당하고 있는 실정에서는 한국은 선진국 대열에 들어갈 수가 없다. 모든 통계인들은 통계의 발전에 모두 다 힘을 모아야 하며, 통계계산분야도 정부에서 주도하고 있는 첨단분야의 일부분임을 알아야 한다. 아직 국내에서 통계패키지를 개발하고자 하는 움직임조차 없다. 조그마한 일부터 차곡차곡 쌓아나가면 후에는 반드시 이루어지리라 본다.

참 고 문 헌

1. 김재주, 조신섭, 김병천(1989). 컴퓨터를 이용한 통계학, 경문사.
2. Kennedy, W.J. and Gentle, J.E.(1980). *Statistical Computing*, Marcel Dekker.