한국통계학회논문집
8권 기념호, 2001
pp. 67-77

생물통계의 과거 · 현재 · 미래

정형기[1], 송혜향[2], 한원식[3], 김병수[4], 이재원[5]

1. 의학분야

1.1. 임상의학분야

국내에서 실시하는 임상시험에는 지금까지는 국내 신약 개발을 위한 3단계의 임상시험과 외국에서 허가되어 2개국이상에서 3년이상 판매하지 않은 신약의 국내도입 과정에서 요구하는 임상시험으로 구분되어졌다. 국내 임상시험에는 주로 외국에서 개발된 신약을 유효성 및 안전성을 확인하는 과정을 지니고 있으므로 피험자수 산정에 가끔 문제가 있었다. 오래전 즉 일본이 ICH guideline을 채택하기전의 피험자수 산정방식을 도입하여 외국에서 개발된 수입약은 30매, 그리고 국내신약은 90매를 해당 질병의 유효성 정도에 관계없이 정하였으나, 그 후 다소 수정하여 예를들면, 두 약군간의 유효율의 차이가 10%내외였어도 무조건 20%차이로 가정하여 피험자수를 산정하여 결과가 통계적으로 유효한 차이가 나지 않아도 즉 점추정치만 차이가 나도 인정하거나 검정력의 개념없이 신뢰구간의 역수로 피험자수를 산정하여 임상시험 결과 산정된 유효성의 95% 신뢰구간이 임상시험계획서상에 언급된 유효율를 포함하면 시험약의 유효성을 인정하는 경향이었다. 이는 피험자수가 증가하면 신뢰구간이 좁아지거나 시험약의 유효율이 예상보다 훨씬 높게 나오면 기존에 언급한 유효율을 포함하지 못하게되는 모순이 있었다.

98년도부터 중앙약사심의위원회 3개분위에 최소 1명씩의 생물통계학교수가 참여하여 각 임상시험에 적절한 피험자수를 산정하도록 자문하였으나 때로는 과다한 피험자수가 산정되어 국내 임상시험 현실상 어려움이 많았다.

그러므로 비교적 객관적 자료가 있는 질환의 임상시험일 경우 과거 대조군 즉 위약군의 유효성보다 비교우위를 보이는 단일 임상으로 실시하여 피험자수를 대폭 줄이거나 비교임상시험의 경우도 비교우위를 보일수 있어도 치료적 동등성 내지는 비열등성 시험 등으로 실시하는 경우가 종종 있었다. 이는 세계시장을 대상으로 하는 개발국과 달리 국내 임상은 국내 시장만을 대상으로 하는 점과 국내의 임상시험에 대한 환자나 그 가족들의 인식 때문에 피험자 모집의 어려운 점도 그 원인이라 할 수 있다.

앞으로는 외국에서 개발된 신약으로서 2개국이상에서 3년간 사용의 조건을 충족 시키는 신약의 경우도 가교자료 예를들면 코케시안과 한국인간의 인종적 특성에 유효성과 안전성면에서 차이가 없다는 자료를 보여야한다. 이에는 약물의 특성상 인종적요인에 영향을 받지 않는 경우와 예민하

1) 서경대학교 수리정보통계학부
2) 가톨릭대학교 의과대학 통계학과
3) 농촌진흥청 농업경영관실
4) 연세대학교 응용통계학과
5) 고려대학교 통계학과

게 영향을 받는 경우로 구분된다. 가교연구에는 AUC나 $t_{\frac{1}{2}}$와 같은 PK를 보는 약동학 연구, 항고혈압제의 경우 혈압대신에 ACE의 억제 효과를 보는 약력학 연구, 상용용량에 있어서 인종간 차이가 보일것으로 예상되는 약물의 경우 용량반응 연구, 타당성이 입증된 대리결과 변수를 사용한 연구, 임상적 결과변수를 사용한 단기간의 안전성캶유효성 확증시험 및 위에서 주로 언급한 현재 실시하고 있는 제3상 임상시험으로 대별 되겠다.

위와 같은 흐름에 따라 앞으로 국내임상건수는 많아지면서 임상분야의 통계수준도 향상되리라 본다. 우선 국내 CRO도 현재 5개나 되며 각 CRO에 통계 전문가를 1인내지는 2인을 두고 있다. 단일 제약회사에서는 년간 임상건수가 많지않으므로 자체적으로 임상팀을 구성하지 않거나 CRA(clinical research association)를 구하기 힘드므로 주로 CRO에 의뢰하는 경향이 증가하고 있다. 이러한 추세로 보아 향후 10년간의 임상의학통계분야는 발전 하리라고 예상된다.

위와같은 추세에도 불구하고 우리가 간과해서는 안될점은 제약선진국의 통상압력에 힘입어 가교시험을 면제받으려고 국내제약사 및 다국적 제약사들의 노력과 이에 편승하여 가교시험 면제 방법을 통계적측면의 고려없이 모색하려 한다는 점이다. 비록 가교시험을 실시하는 경우도 개발국 본사에서 자료를 수합하여 직접 통계처리를 하는 경우가 증가하고 있다. 또한 이전에는 개발국에서 제3상 임상시험이 종료되어 허가가 나야 국내에서 임상시험이 가능하였으나 지금은 개발국에서 허가가 나기 이전에 동시에 다국적 임상시험의 일환으로 국내에서도 소규모로 실시하는바 그 결과를 개발국 본사에서 직접 통계처리를 실시하여 가교자료로 사용하여 제출하고 있는 실정이다. 우리는 이와같은 점을 예의주시하여 앞으로의 임상의학통계분야의 발전을 도모해야한다.

1.2. 병원통계상담분야

통계상담에 있어 매우 두드러진 변화를 볼 수 있었던 분야 중 하나가 의료인을 대상으로 한 병원에서의 통계상담 영역이라 할 수 있다. 우선 폭등하는 상담의 요구를 잘 해결하는 것이 병원 통계상담의 성공 열쇠가 되는데, 이 중 가장 중요한 것은 상담은 상담자와 의뢰자가 일대 일로 부딪쳐서 진행되는 것이 보통이므로 많은 시간과 인력의 해결이 관건이 된다. 병원 통계상담의 주된 인력 원동력은 교수진과 석사 수준의 생물통계연구자(biostatistician)이며 지난 10년간 이러한 인력은 대학원 과정에 생물통계학과(Department of Biostatistics)로 독립된 과가 있거나 또는 보건학과에 속하면서도 생물통계학전공으로 인력을 배출하여 병원 통계상담을 해결하는 큰 변화를 보게 되었다. 이러한 인력배출의 변천은 의료 분야와 통계 분야 모두에 상당히 고무적인 일이다. 상담과 더불어 의료인을 대상으로 한 지속적인 통계 특강도 상담의 일부라 하겠으며 이러한 특강에서 의학논문에 실린 통계의 해석을 도와주며 통계의 이해를 높이며 통계상담의 필요성에 대한 인식도 높이게 된다. 적용되는 통계 방법론 면에서도 매우 차이가 있는 다양한 의학분야의 요구를 해결하는 것은 의학 통계인에게 결코 쉽지 않은 일이지만 이러한 요청에 충실히 대응하기 위해 지속적으로 배워가고 연구하는 과정은 의학통계의 발전에 큰 부분을 차지하고 있음이 사실이다.

상담하는 통계분석법에서도 많은 변화가 있었다. 첫째로 두드러진 변화는 연구계획과 관련된 상담이다. 표본수의 결정에 대한 상담은 80년대에는 년 몇 건에 불과했으나 90년대에 들어서서 매우 빈번하며 더불어 연구계획의 선정에 대한 문의, 및 실제로 연구를 함께 계획하고 진행을 주관하게 되는 역할이 증대되었다. 이러한 상담 및 도움은 자연스럽게 공동연구로 이루어지는 임상시험의 연결선이기도 하다. 의학의 여러 분야에서 임상시험에 참여하는 현실은 뚜렷한 의학 통계인 역할

에 대한 인식과 더불어 의학 통계인에게는 매우 고무적인 일이다. 진료 변화를 모색하기 위한 임상시험, 예를 들어서 중환자실에서 인공호흡기를 부착하고 있는 급성 호흡부전(acute respiratory failure) 환자에게 효과적인 가스교환을 위한 체위변경을 알아내기 위해 실시한 라틴방격을 이용한 임상시험은 임상의를 비롯한 제의료인들의 통계적 방법에 대한 이해가 없이는 결코 실행될 수 없는 일이며, 그 결과가 임상에 직접 반영되는 것을 지켜보는 것은 의학 통계인에게 보람된 일이다.

둘째로 컴퓨터의 보편화로 인해 대규모 자료의 관리 및 장기간 관측자료(longitudinal data)의 관리에 요구되는 기술 및 프로그램의 교육을 요청 받고 있다. 다시 말하면 자료관리(data management)에 대한 상담이다. 특히 의학의 제분야 학회(예를 들어서 신장학회, 호흡기학회, 노인병학회, 등)에서 실시하는 등록부(registry) 제도에 의한 전국 환자자료의 수집이 지속적으로 실시되는 경우도 이에 속한다. 예를 들어서 Excel 수준으로 해결되던 몇 년전만 해도 요사이는 Excess로 더욱 방대한 자료를 다루는 프로그램을 요구하게 되었고 요사이는 또 다른 요구가 대두되고 있다. 이러한 자료수집 결과로서 통계분석에 있어서의 변천도 따라왔다.

셋째로 통계분석법의 변천이다. 상담의뢰자가 요구하는 통계분석법일 수도 있겠고 상담의뢰 자료에 가장 적합하다고 판단하여 상담자가 선택한 통계분석법이기도 하다. 「생물통계의 현황(1981-1990)」(신한풍 등, 1991) 에서 보면 장기간 관측자료에 대한 비교적 단순한 분석법인 반복측정 분산분석법, 생존분석법 또는 로지스틱 회귀분석법으로 분석된 경우가 그리 많지 않았으나 지난 10년간 이러한 분석법이 혼하게 사용되고 있다. 특히 이산형, 순위형 또는 연속형의 관련된(correlated) 반복측정 자료의 분석법, 랜덤효과 모형, mixed 효과모형 등을 이용한 분석법의 요구도 많아졌다. 또한 임상시험에서 수집되는 삶의 질 변수에 대한 분석, 환자의 noncompliance 문제도 새롭게 대두되고 있다.

넷째로 여러 연구결과를 통합한 메타분석의 요구가 또 다른 변천이다. 나날이 축적되어 가는 연구결과를 통계적인 방법으로 요약하는 분석법인 메타분석법에 대한 요구는 자료의 전산화로 인한 수량적 연구결과의 축적으로 인해 자연스러운 결과라 할 수 있다.

2. 생물정보학분야

생물학 분야에서 단백질의 3차원 구조를 밝히는 구조생물학은 전산 방법을 많이 사용하고 이 분야를 흔히 전산생물학(computational biology)이라고 부르고 있다. 지난 10여년 동안 인간 게놈(genome)의 염기 서열 해독이 이루어지면서 전산생물학은 수학 및 통계학과 학제간 병합을 이루어 생물정보학(bioinformatics)이라는 새로운 학문분야로 대두되었다. 여기에서는 생물정보학에서 통계학이 가장 활발하게 응용되고 있는 microarray분야에 대하여 고찰하기로 한다.

2.1. 인간 DNA 염기서열의 해독과 그 이후

인간 genome의 DNA 염기서열을 해독하고자 하는 Human Genome Project가 미국과 유럽 일부 국가들의 협력으로 1990년 시작된 이후, Celera Genomics 사의 가세로 그 연구 속도에 가속도가 붙게 되었다. 그 결과, 예정보다 2년 이상 빠른 지난 2월, 29억여 개의 염기서열(99%)이 확인된 유전자 지도가 발표되어 과학자들뿐만 아니라 일반인들도 유전자, 특히 질병 관련 유전자에 대하여 관심을 갖게 되었다 (Venter et al., 2001). 하지만 실제 사람의 유전자수는 예측한 바와 달리 35,000-40,000 개로 알려지고 이들 중 오직 1.1% 만이 실제 단백을 coding하는 exon (intron

24%, intergenic DNA 75%)임이 밝혀지게 되었다. 이들 유전자 중 기능을 알고 있는 유전자는 일만 개 이하이며, 그 기능의 복합성으로 인해, 유전자 해독이 끝나면 모든 유전자 관련 질병의 진단과 치료 및 생명체의 생성, 생장 및 노화에 이르기까지의 비밀이 풀릴 수 있으리라는 기대와는 달리 더욱 많은 유전자의 기능적 연구가 진행되어야 함을 시사하고 있다. 결과적으로, 21세기에는 genome, 즉 유전체의 1차 구조에 대한 이해를 토대로 하여 각각의 유전자 수준을 넘어 인간의 유전체 규모에서 유전자와 생명현상과의 관계를 규명하고 유전자의 이상이 질병의 발생과 현상 발현에 어떠한 영향을 미치는 지를 규명하는 일이 가능하게 되었다. 이러한 일을 담당하는 새로운 패러다임으로서 microarray를 들 수 있다.

2.2. DNA Microarray와 응용

DNA chip/microarray란 작은 glass나 membrane 위에 수천, 수만 개의 유전자를 얹어놓고 검사하고자 하는 시료에서 추출해 낸 RNA의 발현 정도를 한 번의 실험으로 조사할 수 있는 high throughput screening 방법이다. 1995년 미국 스탠포드 대학의 Pat Brown 등에 의하여 개발된 이 DNA chip은 마이크로프로세서 닮은꼴의 조그만 면에 여러 유전자의 cDNA를 붙여 놓은 것이다. 여기에 형광물질로 표지된 시료를 가한 후에 보합반응 (hybridization)을 시킴으로서 DNA나 RNA의 특징적인 위치(돌)를 알 수 있을 뿐 아니라, 그 위치의 염기순서와 신호의 强度로부터 많은 보합 관련 정보를 얻을 수 있다(Marton et al., 1998; Ross et al., 2000; Scherf et al., 2000). 반도체 기술과 화학기술의 발달, 그리고 생물정보학의 발달로 방대한 연구결과의 분석이 가능하게 되었으며, 이를 통하여 질병의 진단과 치료 개념의 변화가 나타나고 있다. 이와 같은 DNA chip의 장점으로는 동일한 chip의 다량 제작이 가능하며 수백 개 이상의 유전자를 동시에 검사 가능하게 되어 기존의 방법보다 수백 배 이상의 시간과 비용을 절감할 수 있으며, 극히 미량의 DNA로도 chip 제작이 가능한 점등을 들 수 있다.

Microarray는 붙이는 유전물질의 종류에 따라 크게 cDNA microarray와 oligonucleotide array로 나눌 수 있다. 1994년 Affymetrix 사에서 처음 제작한 oligonucleotide array는 1.28 ㎠의 사각형 칩에 photolithography를 이용하여 oligonucleotide를 20여 개씩 붙여나가는 방법을 사용하였다 (Golub et al., 1999). 이 microarray는 유전자의 발현양상뿐 아니라 유전자 sequencing, 돌연변이 및 SNP(single nucleotide polymorphism)까지 발굴 가능한 장점이 있으나, chip 제작 방법상의 어려움, 고가의 chip, hybridization과 scanning 단계를 위한 특수 기기의 필요로 인한 실험비용의 증가 등 경제적인 어려움이 많이 대두되고 있다. 이를 보완하여 발달된 반도체 기술과 화학기술을 이용하여 유리 슬라이드 위에 cDNA를 点積한 cDNA microarray가 개발되었다.

즉, DNA chip을 이용함에 따라 세포 및 조직의 생리학적, 또는 병리학적 변화에 따라 유전자들의 패턴이 어떻게 변하는지 종합적으로 파악할 수 있게 되었다 (Hilsenbeck et al., 1999; Alizadeh et al., 2000; Perou et al., 2000). 또한 생리학적 변화 뿐 아니라 외부적 처치 및 자극에 따른 반응을 동시에 관찰하면서 미지의 유전자들의 역할을 추정하고 궁극적으로 개개의 유전자들의 기능을 밝힐 수 있게 되었다 (Atul et al., 2000). 특정 유전자의 이상 또는 발현 변화는 단지 그 유전자의 변화만으로 끝나는 것은 아니며, 또 다른 여러 유전자의 발현 변화를 유도함으로써 최종적으로 특정 형질의 발현을 유도하게 된다. 다시 말하면, DNA chip은 특정 유전자의 이상이나 발현 변화로부터 나타나는 다차원적인 pathway에 의한 유전체 활성의 총체적인 변화를 규명할 수 있는 방법이다. 이러한 연구를 통하여 각종 질병의 진단 및 신약 개발에 필요한 치료 목표와 정보를 얻을

수 있다. 즉, 각종 질환에서 질병의 발생 단계에 따른 유전자 발현의 변화를 관찰하고 각 질병 고유의 특징적인 유전자 발현 패턴을 이해함으로써 진단, 치료의 효과적인 대상 유전자를 규명할 수 있게 되어 DNA chip이 21세기 생명공학의 시대를 주도할 것으로 예견하고 있다 (Maton et al., 1998; Golub et al., 1999; Bittner et al., 2000).

2.3. DNA Microarray 자료와 통계분석의 필요성

대부분의 cDNA microarray 실험의 기본 목적은 군간 발현의 차이를 나타내는 유전자들을 검색하는 것이다. 따라서 단순하게 몇 배수원칙(fold change rule)을 적용하는 것보다는 위양성 및 위음성 확률을 최소화하는 적정한 통계 분석 모형을 구성하여 유의적 차이를 보이는 유전자들을 찾는 것이 중요한 생물학적 변화와 우연에 의한 변화를 구별하여 주는 방법이라고 하겠다 (Tanaka et al., 2000, Fig. 2). 뿐만 아니라 일반적으로 새로 개발되는 assay는 널리 보급되기 전에 재생가능성과 실험의 변동 요인등에 대한 철저한 규명이 따르게 되는데 이러한 연구는 ANOVA등과 같은 통계적 분석이 수반된다.

두 염료 (Red, Green)의 색의 강도를 같게 하여 주는 표준화를 실시하고 그 결과로 얻어진 (R, G)를 사용하여 우선 microarray 실험의 재생가능성을 검토하는 통계적 분석 모형을 구성하고 적정 반복 실험 회수를 현실적으로 가능한 1-3회 사이에서 결정하여야 한다. (Lee et al., 2000; Kerr et al., 2001). 또한 실험의 성격상 먼지등으로 인하여 생기는 결측치에 대한 대치(imputation) 문제도 효율적인 방법으로 해결이 되어야 한다. (Troyanskaya et al., 2001). 이상의 처리가 모두 이루어지면 본격적인 자료분석을 시도할 수 있게 된다. Microarray 실험은 본 연구자들 뿐 아니라 대부분의 국내 실험실에서도 경험이 일천하므로 엄격한 실험 설계하에 동 실험의 변동요인을 밝히는 작업도 병행되어야 하며, 이렇게 찾아진 변동요인은 추후 자료분석의 근간을 구성하게 된다.

Microarray 실험은 실험의 속성상 대량의 자료를 생성하게 된다. 가령 probe로 사용된 유전자가 6000개이고 표본이 30개(대조군 15개, 처리군 15개)일 경우 6000*30=180,000개의 관찰치가 産産되며, cDNA Microarray와 약간 다른 실험 설계하에서 구성된 Oligonucleotide array를 사용하는 경우는 각 유전자마다 20개의 perfect match, mismatch 쌍 (pm, ms)이 있게되어 20*2*30*6000=7,200,000개의 자료가 분석 대상이 된다 (Efron et al., 2000). 많은 양의 자료를 효율적으로 축약하여 가령 두 군간의 유의적 차이를 나타내는 유전자군을 도출하는 일, 그리고 유전자군간의 상호작용을 규명하는 일은 바로 통계학의 고유한 영역이 된다.

이제 21세기의 생물학자들은 자료를 수집하는 일보다는 자료를 효율적으로 분석하는 일을 더 중요시하게 되었고 이러한 맥락에서 통계학 및 전산과학과의 학제간 연구는 생물정보학이라는 새로운 학문을 탄생시켰으며, 이 생물정보학이 21세기 생명과학기술(BT)의 기초 학문이 되고 있다. (Cech, 2001). 특히 생물정보학의 중요한 부분을 담당하고 있는 Microarray 관련 통계이론들은 bootstrapping (Kerr and Churchill, 2000), missing data imputation (Troyanskaya et al., 2001), 다중 검정으로 인한 P값 보정 절차 (Dudoit et al., 2000b), 군집분석 (Eisen et al., 1998), 주성분분석 (Hilselbeck et al., 2000, Hastie et al., 2000), empricial Bayesian 이론, 정규분포의 혼합모형 (Lee et al., 2000), 회귀분석, Cox의 비례위해모형과 군집분석을 혼합하는 기법 (Hastie et al., 2001), 실험계획법 (Kerr and Churchill, 2001a) 등 거의 통계학 이론의 백화점이라고 할만큼 다양한 이론이 적용되고 있다. 20세기의 통계학은 농학의 발달과 함께 토지에 대한 수확량을 늘리기 위하여 땅을 대상으로 실험계획을 시행하면서 많은 발전을 이루었고, 이 조그만 유리판 위에서 시

행되는 21세기의 실험계획도 기존의 실험계획법으로부터 많은 것을 배울 수 있을 것이다. (Kerr and Churchill, 2001a).

3. 농업연구분야

3.1. 농업연구에서의 통계분석의 현황

우리 나라의 농업과학기술은 근대이후 끊임없는 발전을 거듭하여 왔다. 새로 개발된 기술의 성과가 실제 농업 연구현장에서 보다 정확하게 적용되기 위해서는 연구결과의 신뢰성과 정밀성이 무엇보다 중요하며 여기에는 통계적 방법이 매우 중요한 역할을 한다고 하겠다. 최근에 우리의 농업시험연구는 새로운 첨단기술을 비롯한 연구분야의 전문화 및 다양화로 연구방법이 고도화되고 있는데 통계적 방법은 연구를 수행하는 과정에서 시험의 합리적 설계와 정확한 결과분석의 기본적인 과정이라 할 수 있다. 따라서 현재 수행하고 있는 농업시험 연구의 설계와 통계분석을 보다 과학적으로 분석한다면 한층 신뢰성 있는 연구결과의 도출이 가능해지고 이것은 바로 우리 농업시험 연구사업의 효율성을 높이는 것이라고 하겠다. 이에 농촌진흥청에서는 연구사업의 증진을 위한 노력의 일환으로 작물, 원예, 축산 분야 등에서 각 분야에 적합한 통계사업을 추진하고 있으며, 여러 가지 다양한 통계적 기법들이 사용되고 있다. 통계기법 중 실험계획법에 관한 분야는 많은 부분이 농사시험에서 발견된 것과 같이 여러 종류의 실험계획법이 가장 다양하게 사용되는 것이 바로 농업분야이고, 이러한 분야는 농촌진흥청에 통계적 분석을 위한 소프트웨어 및 컴퓨터가 보급된 이래 꾸준히 유용하게 널리 사용되고 있는 분야라 볼 수 있다. 한 예로 97년도 농촌진흥청 205개 연구사업을 대상으로 조사한 자료에서 연구사업에 사용된 실험계획 및 통계분석현황을 살펴보면 [표 1]과 같다.

[표 1]. 농업시험에서의 시험구 배치 사용 빈도

시험구배치	빈도
난괴법 3반복	50
난괴법 4반복	1
분할구 배치 3반복	8
완전임의배치 3반복	18
완전임의배치 5반복	6
단구제	31
파종시기별 strip 배치	1
사각 폿트 시험	2
기타(조사연구)	88
합계	102

[표 1]에서 보면 시험구 배치에 있어서 난괴법과 임의배치법등이 가장 많이 사용된 것으로 나타났다. 그러나 총 205과제중에서 시험구 배치를 언급한 경우는 102과제로 전체의 50%에 해당된다. 사용된 통계분석 방법도 보면, 다중비교나 빈도분석, 분산분석, 인자분석등이 차지하는 비율이 약 40% 정도로 단순한 평균비교검정이나 빈도분석 상관분석등이 대부분이었다. 이처럼 시험구배치에서 완전확률화 계획법, 난괴법, 분할구 계획법, 격자계획법등 한정된 방법이나 단순한 분석기법을 연구자들이 선호하고 있는데, 이는 실험설계나 분석과정에서의 어려움 등이 있기 때문으로 해석된다. 따라서 다양한 통계사업을 추진함으로써 이러한 어려움을 극복하기 위한 노력을 계속하고 있다.

3.2. 최근 농업분야에 적용되고 있는 통계적 방법 및 앞으로의 현황

최근 농촌진흥청에서는 새로운 통계적 방법들을 다양한 분야에서 접목이 시도되고 있다. 예를 들어 분산분석 모형 및 베이지안 이론 또는 몬테칼로 시뮬레이션등의 통계적 방법을 이용한 유전·육종 분야의 통계분석에 관한 연구 및 확률론 및 MCMC 방법등을 이용한 농업생물자원 정보분석에 관한 연구에 많은 관심을 가지고 연구를 추진하고 있다. 가축통계분석에서 사용되고 있는 통계적 방법을 몇 가지 소개하면 다음과 같다

① 가축 육종을 위한 분산성분 추정

분산 및 공분산 성분의 추정은 가축 육종에 있어 1) 선발지수 식의 설정 2) 모형에서 BLUP을 얻기 위하여, 3) 유전력, 유전상관, 표현형상관 및 환경상관 등의 추정, 4) 육종계획의 수립, 5) 양적 형질에 대한 유전적 작용의 설명 등에 널리 이용되어진다(Henderson,1986). 특히, 가축의 유전능력을 추정하기 위한 혼합모형을 적용하고 할 때 육종가를 포함하고 있는 임의효과부분에서 분산성분의 이용은 필수적이다.

분산분석(ANOVA)에 근거하는 분산성분의 추정방법은 지난 50여년 동안, Henderson method 1,2,3, MINQUE, MIVQUE, ML, REML, Method R 및 깁스 샘플링 등으로 추정 알고리즘이 발전되어 왔으며, 이와 더불어 평가에 이용되는 가축의 수도 과거 수백마리에 불과하던 것이 현재는 수만, 수십만 마리가 동시에 평가되어 지며, 이에 따라 혼합모형의 방정식의 수도 이에 비례하여 증가하기 때문에 추정값의 정확도뿐만 아니라 계산비용 및 시간을 줄이는 것도 중요한 문제로 대두된다.

1970년대까지 분산성분을 추정하는 기본적인 방법은 각 방법별로 약간의 차이는 있지만 정규분포하에서 관측치 벡터의 2차 형식을 정의하고 이에 대한 기대값을 계산하여 양 항들 부동항으로 미지의 모수를 추정하는 거이 일반적이었다. 그러나 이후에는 Likelihood를 이용하여 추정하는 방법들이 일반화되었다. 그러나 Likelihood 방법들은 계산식이 복잡하고 분석에 많은 반복 추정을 필요로 하여 계산 시간과 비용을 절약할 수 있는 깁스 샘플링을 이용한 Bayesian 접근 방법이나 Method R과 같은 간편한 추정법들이 개발되었다. 이러한 통계적 방법들이 농업축산분야에서 유용하게 사용되고 있다.

② 가축육종분야의 모의실험

시뮬레이션에 사용되는 모형은 현실 실존물의 대응물로 실존물의 주요 특성을 포함하여야 한다. 그러나 실험 목적과 관계없는 특성은 포함시킬 필요는 없다.

• 단형질 자료의 생성 : 형질의 평균값이 M이고, 분산이 V인 자료를 생성한다면 다음과 같다.

$$Y_i = M + R_i\sqrt{V},\ \ i = 1, \cdots, n$$

Y_i 는 한 개체의 단일 형질 측정값이며, M은 표현형 평균, V는 표현형 분산, R_i는 표준정규분포변수(난수)이다. 표현형 분산 $V(P) = V(G) + V(E)$이고, $COV(G, E) = 0$ 으로 가정한다. 이러한 방법으로 생성된 자료는 평균과 분산만 고려한다면 현실에서 얻어진 자료와 구별할 수 없다. 그러나 이 자료는 정규분포한다는 가정하에 만들어진 것으로 다른 통계량으로 비교하면 다를 수 있다.

단형질 자료 생성방법으로 여려 형질의 자료를 생성하더라도 형질간의 연관성이 없으므로 모의실험 자료로 사용하기에 부적합하다

• 다형질 자료의 생성 : 다형질 자료를 생성하기 위해서는 각 형질의 평균과 분산, 그리고 각 형질들간의 공분산이 필요하다. 공분산에 의해 각 형질간의 연관성이 결정된다. 다형질 자료를 생성하기 위해서는 각 형질의 분산·공분산 행렬을 촐레스키 분해(Choleskey decomposition)한 후 정규분포 난수와 곱하여 변이가 주어진 새로운 변이를 생성하게 된다.

③ 농업생물정보 DB화 및 분석

최근 BT를 통한 농업연구분야에서의 연구체제는 기존 연구체계와의 통합적 추진체계로 설정되어가고 있다. 이는 여러 가지 실험환경 및 연건의 변화, 소비자의 다양한 기호충족 및 저 비용 고효율 농업산업의 자구책 마련의 계기라 여겨진다. 또한 단위기술 및 단위학문영역에서 생산물 중심의 통합적 연구체제로의 전환이 이루어지고 있다. BIO Informatics(수리통계학, 통계육종학, 컴퓨터공학, 분자생물학)이나 BIO Diversity(분자유전학, 통계육종학, 집단유전학)와 같은 인접 최첨단 기술체계 연계활용이 결과도출 및 산업적 용의 성공여부 관건의 인식되고 있다. 농촌진흥청의 생명과학 산업도 농업과학기술원을 중심으로 활발한 활동을 하고 있다.

• 농업생물 게놈(genome)연구 및 자원관리 정보축적

농업생물분야의 지놈 연구는 1994년 과기부의 선도기술개발 사업으로 농업과학기술에서 국가적인 벼 게놈연구 사업(Korea Rice Geome Research Program)을 수행하면서 본격적으로 시작되었다. 여기서는 분자유전자지도 작성, 대량 cDNA 염기서열 분석, 유전자지도를 기초로 한 유용 유전자 클로닝, 형질전환 및 분자육종 실용화 기술개발에 중점을 두고 추진하였다. 분자유전자 지도는 통일계 "밀양 23호와 일반계 "기호벼"를 교배하여 164계통을 육성하여 작성하였다.

또한 생명공학 원천기반 기술 개발을 촉진하기 위하여 생물자원 및 유전체 분석 정보를 체계적으로 제공할 수 있는 정보관리 시스템 구축하고 이를 위하여 주요작물 및 미생물의 유전체 분석 정보, 분자유전자 지도정보 및 유전자변형작물 개발 및 안전성 관련 정보를 데이터베이스화하여 웹사이트(http://biogen.nisat.go.kr)에 공개하였다.

3.3. 농업연구에 있어서 통계기법의 다양화 및 제안점

1990년대 이후 지난 10여년 동안 농업연구분야에서도 다양한 통계적 접근이 시도되었고, 그 결과 상당한 진전이 있었던 것으로 나타났다(작물학회지, 육종학회지등을 포함한 농업관련 학회지에

게재된 논문에서 조사된 통계적 방법). 과거와는 달리 프로그램상의 통계적 기법을 이용하는 것에 그치지 않고, 요인분석, 다차원 척도법, 판별분석등과 같은 다변량 기법의 효과적이 활용이나 농업기상관측을 위한 통계적 예측모형의 적용과 같은 고급 통계이론을 농업연구분야에 적절하게 응용하는 사례도 상당한 증가 추세를 보이고 있다. 그러나, 농업연구자들의 통계적 방법의 적용에는 여러 가지 어려운 점이 많을뿐더러, 그동안 통계 전문가들의 농업생물분야에 대한 관심이 미비하다는 것이 한가지 흠이라 생각되어진다. 농업연구는 생명과학 연구의 핵심분야로서 농업생명에 관한 단독 연구로서는 더 이상의 발전은 없을 것이며, 앞으로는 다양한 분야와의 접목이 필요하며 특히, 통계 전문가들의 농업생물통계에 관한 실용적 접근이 절실하게 요구된다고 볼 수 있다.

참고문헌

[1] 농촌진흥청 (2001). 가축통계유전분석에 관한 통계 세미나.

[2] 농촌진흥청 (2001). 생명공학과 농업환경 연구를 위한 정보기술 이용.

[3] 신한풍,송혜양,김병수,이종협,한원식 (1991). 생물통계의 현황(1981-1990). 통계학연구, 62-91.

[4] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A,Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Scherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Bryd JC, Botstein D, Brown PO, Staudt LM. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature, 403:503-511.

[5] Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Ranmacher M, Simon R, Yakhini Z, Ben-Dor A, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Gillaners E, Leja D, Dietrich K, Berens M, Alberts D, Sondak V, Hayward N, Trent J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature, 406:536-540.

[6] Cech TR. (2001). The $13-billion man, Scientific Amer. January 2001:22-23.

[7] Dudoit S, Yang YH, Callow MJ, Speed TP. (2000b). Statistical methods for differentially expressed genes in replicated cDNA microarray experiments. submitted to J. Amer. Statist. Assoc., www.stat.berkeley.edu/users/terry/zarray/Html/papersindex.html.

[8] Eisen MB, Spellman PT, Brown PO, Botstein D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., 95:14863-14868. Efron B, Tibshirani R, Goss V, Chu G. (2000). Microarrays and their use in a comparative experiment. Tech Report (Oct.3, 2000). Dept. of Statistics, Stanford University.

[9] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. (1999). Molecular classification of cancer: Class discovery and classprediction by gene expression monitoring. Science, 286:531-537.

[10] Hastie T, Tibshirani R, Botstein D, Brown P. (2001). Supervised harvesting of expression trees. Genome Biology 2:research0003.1-0003.12.

[11] Hastie T, Tibshirani R, Eisen M, Alizadeh A, Levy R, Staudt L, Botstein D, Brown P.

(2000). 'Gene shaving' as a method of identifying distinct sets of genes with similar expression patterns. Genome Biology 1:research0003.1-0003.21.

[12] Henderson, C.R. (1984). Applications of linear models in animal breeding. Univ. of Gueiph, Canada.

[13] Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK, Fuqua SAW. (1999). Statistical analysis of array expression data as applied to the problem of tamoxifen resistance, J. Natl Cancer Inst. 91:453-459.

[14] Kerr MK, Afshari CA, Benett L, Bushel P, Martinez J, Walker N, Churchill G. (2001). Statistical analysis of a gene expression microarray experiment with replication. Tech Report, Jackson Lab, www.jax.org/research/churchill/pubs/index.html.

[15] Kerr MK, Churchill GA (2001a). Experimental design for gene expression microarrays, Biostatistics, 2:183-201.

[16] Lee M-LT, Kuo FC, Whitmore GA, Sklar J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive DNA hybridization. Proc. Natl. Acad. Sci. 97:9834-9839.

[17] Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, Bassett DE Jr, Brown PO, Friend SH. (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat. Med. 4(11):1293-1301.

[18] Perou CM, Sorlie T, Eisen MB, Van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Bostein D. (2000). Molecular portraits of human breast tumours, Nature. 406(6797):747-752.

[19] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Bostein D, Brown PO. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet. 24(3):227-235.

[20] Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. (2000). A gene expression database for the molecular pharmacology of cancer, Nat. Genet. 24(3):236-244.

[21] Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraya R, Doi H, Wood lll WH, Becker KG, Ko MSH. (2000). Genome-wide expression profiling of mid-gestation placentaand embryo using a 15,000 mouse developmental cDNA microarray. Proc. Natl. Acad. Sci. 97:9127-9132.

[22] Troyankaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001). Missing value estimation methods for DNA microarrays, Bioinformatics, 17:520-525.

[23] Venter CJ, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman

JR, Zhang, Q, Kodira CD, Zheng XH, Chen L, Skupsk M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, et al. (2001). The sequence of the human genome. Science 291:1304-1351.

[24] Yang YH, Buckley MJ, Dudoit S, Speed TP. (2000a). Comparison of methods for image analysis on cDNA microrarray data. Tech Report, #581, Dept. of Statistics, Univ. of California at Berkeley.

[25] Yang YH, Dudoit S, Luu P, Speed T. (2000b). Normalization for cDNA microarray data. Tech Report #589, Dept of Statistics, Univ of California at Berkeley, www.stat.Berkeley.EDU/users/terry/zarray/Html/ papersindex.html.