

통계학적 사고의 역사

한국통계학회

2018. 봄

조재근 (경성대학교)

1. 들어가는 말

의학의 역사를 소개하는 번역서를 펼쳤다가 맨 앞에서 이런 부분을 만난 적이 있습니다.

“1928년에 처음 의과대학에서 공부를 시작하였을 때 우리는 그 시대의 의학이 이전 시대 의사들의 경험적 방법과는 크게 다른 것으로서 진정한 과학이라고 믿었다. ... 그런데 금세 그런 공부가 남아버리고 한때 자신만만하던 우리가 시골 의사 취급을 받고 있더라.”

(<의사들의 생각: 그 역사적 흐름>, 킹, L. 지음, 이흥규 옮김, 고려의학, 1994, p. 3).

그 책을 쓴 이는 “의학지식의 반감기가 5년밖에 안된다는 말도 있던데 그렇다면 지금 새로운 의학도 겨우 10년만 지나면 과학적인 의학이 될 수 없겠다”며 씩씩해 하였습니다. 그런 평가가 맞다면 비록 제가 전혀 모르는 세계입니다만 오늘날의 의학연구자들, 의사들에게는 20세기 전반기는 물론 바로 얼마 전인 20세기 말의 의학도 남아빠진 구닥다리 취급을 받는 모양입니다.

통계학은 어떨까요? 저는 대학 3학년이던 1981년에 (원지도 모르고) 배운 <수리통계학>을 지금도 수업에서 잘 써먹고 있습니다(물론 제가 일하는 학교에는 통계학과 대학원 과정이 없어서 학부 수업만 하기 때문이겠습니다). 그래도 그렇지, 과연 이래도 될까요? ππ

어쨌든 학교에서 익힌 것들의 유효기간만 본다면 지금까지는 통계학과에서 배운 것이 의과대학에서 배운 것보다 훨씬 더 “쓸모 있어” 보입니다. 19세기 영국의 통계를 주름잡았던 William Farr(1807-1883)는 젊었을 때에 의사라는 직업을 버리고 통계공무원으로 평생을 살았는데 오늘날에도 역시 의과대학보다는 통계학과에 가는 편이 더 나은 선택인지 모르겠습니다. ㅎㅎ

통계학의 역사에 조금 관심을 갖고 있는 저는 방금 말씀드린 의학과 통계학의 차이는 그 학문들의 바탕을 이루는 핵심적인 사고방식 같은 것들이 갖는 유효기간의 차이와 관계가 있지 않을까 생각해본 적이 있습니다. 통계학은 어쩌면 학문분야들 가운데 가장 융합적이고 변화무쌍한 분야이기 때문에 끊임없이 새로운 이론과 방법들이 만들어지고 새로운 분야들에 활용되어 왔습니다. 그렇지만 20세기 초반에 나온 K. Pearson, R. A. Fisher, J. Neyman 등의 사고방식과 연구성과는 21세기인 현재까지도 여전히 통계학 교과서에서 핵심적인 자리를 차지하고 있습니다. 사실 Fisher는 1962년에 세상을 떠났고 통계학 분야에서 그가 남긴 중요한 업적들은 1920-30년대에 주로 나왔습니다. 그러므로 의학 분야의 시간감각으로 보면 피셔는

아주 아주 옛날 사람일 것입니다. Neyman 역시 30년도 훨씬 이전인 1981년에 세상을 떠났으니 구닥다리이긴 마찬가지겠습니다.

세월이 더 지나면 어떨까요? 짐작컨대 다음 세기, 즉 22세기에 통계학을 공부하는 사람들도 Fisher의 이름을 여전히 기억하게 될 것 같습니다. 어쩌면 Fisher가 한창 활동하던 1930년대에 태어난 Efron 같은 사람의 이름도 거기에 추가될지 모르겠군요. 그런데 확률이나 데이터, 통계학의 역할, 과학과 통계학의 관계 등에 대한 대가들의 사고와 접근법이 오래 살아남는 이유가 (의학 등의 학문분야들과 비교했을 때) 통계학의 발전속도가 느렸기 때문은 아닐 것입니다. 아마 그 이유는 제가 제목으로 삼은 “통계학적 사고”라 부를 수 있는 것들이 나름 세월의 침식작용을 잘 견뎌내기 때문이라고 보아도 될 것 같습니다.

그렇다면 통계학적 사고란 무엇일까요?

앗! 그리고 보니 허명회 선생님께서 쓰신 책의 제목이 바로 <통계적 사고>¹로군요! 그 책의 머리말을 찾아보니 “통계적 사고의 ‘사고(思考)’는 생각하고 헤아린다는 뜻으로서 ‘사고(事故)’와는 다르다!”라는 재미난 말씀으로 시작하셨어요. 이어서 선생님께서는 통계학은 “사회과학, 자연과학에 속하는 수많은 개별 과학에서 쓰이는 실증적 방법론의 정량적 논리”를 다룬다고 하셨습니다. 선생님이 말씀하신 바에 따르면 ‘통계적 사고’란 실험이나 관측을 통해서 얻은, 우연을 동반한 수량적인 데이터에 대해 논리적으로 접근하는 사고라고 할 수 있을 것 같습니다(적절한 사례를 골라 간결하게 쓰시는 선생님의 글쓰기 스타일은 늘 부러울 따름입니다. 어수선한 글밖에 못 쓰는 저같은 사람한테 특히!).

그런데 통계학자들이 “통계학의 역사는 백년 남짓, 또는 이백년 남짓밖에 되지 않는다”라고들 말하는 것을 보면 그와 같은 논리적 사고가 등장하고 널리 쓰이게 된 과정이 말만큼 쉽지만은 않았던 모양입니다. 영어 단어 statistics를 우리말로 ‘통계’라고도 부르고 ‘통계학’이라고 부르듯이 statistical thinking 역시 통계적 사고, 아니면 통계학적 사고, 이렇게 둘로 옮길 수 있겠습니다. 허명회 선생님께서 말씀하신 ‘수량적 데이터에 대한 논리적 사고’라는 설명에는 데이터로서의 통계와 추론으로서의 통계학이 다 들어있습니다. 여기서 저는 그 둘을 조금 나누어서 살펴볼까합니다. 그래야 역사를 구분하기가 쉬워질 것 같거든요.

아주 거칠게 정리해보자면 19세기가 시작될 시기의 statistical thinking은 “통계학 없는 데이터 중심의 통계적 사고”라고 할 수 있을 것 같습니다. 즉 데이터 자체가(심지어 초기에는 데이터 자체만이) 현상들, 특히 19세기 사람들이 열광했던 사회데이터를 이용하여 사회현상들을 파악하는 가장 과학적인 것으로 받아들여졌습니다. 그렇다면 수학에 바탕을 둔 확률과 통계학 이론은 데이터와 멀었겠지요. 그러다가 19세기의 거의 끝무렵부터 둘이 만났고 그 이후부터

¹ 교우사, 2006.

20세기의 눈부신 성과들이 나오게 됩니다. 즉 19세기는 통계적 사고에서 시작하여 통계학적 사고로 옮겨간 시기라고 할 수 있겠습니다.

여기서 제가 조금 더 거칠게 주장하려는 바는 데이터 중심의 통계적 사고에서 추론 이론 중심의 통계학적 사고로 옮겨가는 변화가 혁명적인 변화임에는 틀림없었지만 돌이켜 보았을 때 그 둘은 공통점이 더 많았다는 것입니다. 제가 19세기와 20세기 통계학을 한 데 묶을 수 있다고 보는 공통점은 개체(개인)가 아닌 집단에 주목하는 사고방식입니다. 사회학이 대표적이고 황승식 교수님의 영역인 역학(epidemiology) 역시 마찬가지겠지요. 처음부터 통계와 가까웠거나 지금도 역시 그러한 이들 학문이 시작된 시기가 바로 19세기였습니다.

그렇다면 지금은 어떨까요? 제가 <통계학적 사고의 역사>라는 자못 거창한 제목으로 발표자료를 준비해볼 생각을 한 것은 21세기에 접어든 이후 여기저기서 듣게 된 새로운 소식들 때문이었습니다. 의학에서는 personalised medicine, 즉 개인별 맞춤의학 이야기를 자주 들을 수 있었습니다. 마케팅 분야에서도 고객들의 개별적인 특성을 잘 파악해야 성공할 수 있는 시대가 되었다고들 합니다. 선거 홍보 분야에서도 오바마가 유권자의 정치 성향을 개별적으로 헤아려서 선거전략을 짜는 마이크로 타게팅이 효과를 거두었다는 이야기를 자주 듣게 되었습니다. 이런 새로운 변화들은 모두 굉장히 많은 데이터를 분석해야 하는 빅데이터시대의 산물들이라고 합니다. 그리고 빅데이터를 분석하는 데에는 각종 통계학 이론과 분석법들이 매우 중요한 역할을 한다고들 합니다.

그런데 잠깐 생각해 보면 이러한 사례들은 모두 집단이 아니라 모두가 서로 다른 개인에 주목하고 있습니다. 그렇다면 이제 다시 시계가 과거로 방향을 바꾸어 통계학에서 집단보다는 개인이 중요한 시대가 된 것일까요? 혹시 바로 지금이 19세기에 나타나서 두어 세기 동안 유지되었던 통계(학)적 사고의 기본틀이 변하고 있는 시기일까요? 어쩌면 한가한데다 너무 거칠어 보일지도 모를 이런 단순한 질문이 과연 질문거리가 되거나 하는지, 이 자리에서 한번 여쭙보고 싶었습니다.

2. 19세기의 통계적 사고: 개인이 아닌 집단을 보라

개인은 불확실성의 원인,
더 많은 데이터를 모아서 불확실성을 줄이자

2.1 국가처럼 보기 또는 통계처럼 보기

제가 좋아하는 책 중에 <국가처럼 보기>라는 책이 있습니다.² 700쪽에 가까운 번역서는 2010년에 나왔고 *Seeing Like a State*라는 제목을 단 영문판 책은 1998년에 나왔습니다. 통계학의 역사를 공부하는 입장에서 이 책을 읽다보면 책의 제목을 *Seeing Like a Statistics*라고 바꿔 읽어도 되겠다는 생각이 들곤 합니다.

사람들은 보통 통계학의 역사에 대해 어떤 질문을 할까요? 아주 가끔씩 저는 “정규분포는 언제 누가 처음 만들었나요?”, “골턴 이전에 상관이라는 개념을 생각한 사람이 있었나요?”, “표본조사 방법을 처음 생각한 사람은 누구인가요?” 등의 질문을 받곤 합니다. 아는

대로 대답을 해드립니다만 사실 저는 그런 질문에는 별로 관심이 없습니다. 위키피디아를 찾아보면 답을 얻을 수 있는 문제들이니까요. 저는 분명한 정답이 있는 그런 문제보다는 “왜 각종 통계조사는 19세기가 되어서야 겨우 시작되었을까?, 그 이전에는 데이터 수집 작업을 못한 것인가, 안 한 것인가?” 등의 질문에 더 관심이 있습니다.

대부분의 통계는 수량화된 데이터를 뜻합니다. 어떤 자연이나 사회, 경제 현상을 측정해서 수량화된 데이터로 만드는 것은 아주 오랫동안 사실상 불가능한 일이었다고 합니다. 중세가 저물고 근대국가가 만들어지면서 국가는 많은 정보를 필요로 했지만 한 나라 안에 있는 지역마다 각종 측정 관행들이 서로 매우 달랐기 때문이었다고 합니다. 그렇게 달랐던 측정의 흔적은 무게나 부피, 거리를 재는 단위에 지금도 남아있습니다. 따라서 국가를 운영하는 측에서는 나라



² 스코트, J. 지음, 전상인 옮김, 에코리브르, 2010.

안에서 쓰이는 도량형이나 측정방법을 통일시키는 것이 중요하고 시급한 일이었겠지요. 더 많은 세금을 거두기 위해서든 인력을 동원하기 위해서든 생체정치(biopolitics)를 위해서든 말입니다.

나중에 결과적으로는 국가에서 정한 통일된 측정법들이 지역 간의 소통과 물자 거래를 원활하게 도와주는 역할을 했겠지만 오랫동안 관행적으로 사용해오던 서로 다른 제도를 하나로 통일시키는 데에는 체계적인 행정시스템을 통한 강압적인 방법이 동원될 수밖에 없었을 것입니다. 흔히 우리가 ‘합리화’한다고 부르는 그 과정은 사람들의 사고와 오랜 사회 관습, 제도 등을 바꿔야하는 일이므로 힘 센 근대국가가 아니면 할 수 없는 일이었겠지요. <국가처럼 보기>(50쪽)에도 실려있는 역사학자 T. Porter의 말을 인용해보겠습니다.

사회가 계량화의 대상이 되려면 개조되어야한다. 사람과 재화의 범주를 정의해야하고 척도는 교환 가능해야 한다. 또 토지와 상품, 그리고 노동과 그 노동의 산물은 동일한 가치를 갖는 금전으로 표현될 수 있도록 고안되어야 한다. 여기에는 베버가 합리화라고 일컬었던 개념이 많이 포함되어 있고, 집중화라는 개념 또한 마찬가지다³.

우리가 지금 객관적이고 보편적이며 비교가능하다고 여기는 각종 표준화된 통계 지표들에는 꽤 오랜 사회적 갈등의 역사가 숨어있다고 보아야할 것 같습니다. 표준화로 인해 아주 많은 문화와 생활방식, 세계관, 사고방식이 사라졌을 것입니다.

국가는 지역별로 다른 도량형을 통일시켰을 뿐 아니라 그런 측정방법의 대상으로부터 개별적인 특성까지 지워야했습니다. 처음 인구조사가 시작되었을 때 사람들은 서로 다른 모든 사람을 똑같이 1이라는 단일한 숫자로 헤아리는 것을 보고 매우 당혹스러워했다고 합니다. 국가처럼 보는 것은 곧 통계가 그러하듯이 동질적이라고 간주되는 개체가 모인 집단을 보는 것, 즉 개별적인 존재의 특성을 지우는 것이었습니다. 이때 개체의 고유한 특성은 집단의 중요한 신호를 가리는 잡음, 불확실성으로 간주되었습니다. 스코트는 이처럼 잡음을 무시하고 신호만 읽으려는 시도를 ‘**가독성(legibility)**’이라는 표현으로 부르고 있습니다. 그 표현을 빌리면 **근대 국가에게 통계란 중앙정부가 읽어낼 수 없었던 각 지역의 도량형, 제도 등을 데이터로 읽어낼 수 있게 만드는 작업**이었다는 말이지요. 이렇게 본다면 통계(학)의 역사가 오롯하게 수학자나 통계학자들만의 것이 되기는 어렵겠습니다.

³ 인용문의 출처는 다음과 같습니다: Porter, Theodore. “Objectivity as Standardization: The **rhetoric of impersonality in measurement statistics** and cost benefit analysis”, *Rethinking Objectivity*, Allan Megill (editor), Duke UP, 1994, p. 201. T. Porter는 통계학자가 아니고 역사학자인데 Stigler의 *The History of Statistics*가 나온 그 해에 스티글러의 책만큼 중요한 통계학사 책을 냈습니다. *The Rise of Statistical Thinking, 1820-1900* (Princeton UP, 1986)이 그것인데 통계(학)적 사고라는 표현을 제목에서 볼 수 있습니다

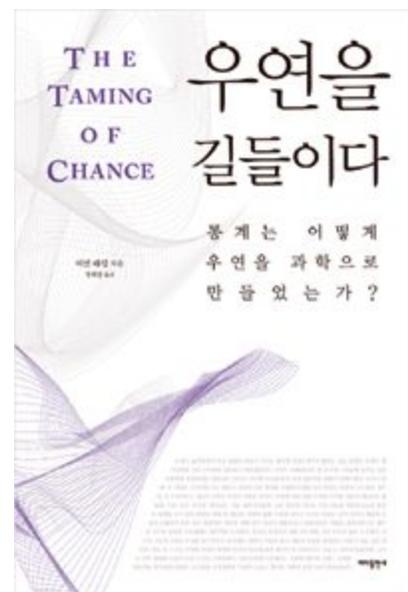
2.2 통계와 개인, 집단을 둘러싼 프랑스의 논쟁

스티글러는 <통계학을 떠받치는 일곱 기둥> 가운데 맨 첫번째 기둥으로 **서로 다른 데이터들의 결합(aggregation, combination of observations)**을 이야기한 바 있습니다. 그는 주로 천문학, 측지학 등에서 나온 자연을 측정된 데이터들의 결합을 다루었습니다. 결합 방법 중에서는 오늘날 우리가 생각하는 **평균**이 대표적인데 사람들이 그런 결합을 받아들이는 것이 쉬운 과정이 아니었습니다. 그래도 이질적인 개체와 집단들을 조사한 경제나 사회 데이터에 비하면 자연 동질적인 자연 데이터를 결합해서 평균을 내면 불확실성을 줄일 수 있다는 것을 납득하는 일은 수월한 편이었습니다. 사회 데이터에서는 어땠을까요?



사회를 이루는 집단과 개인의 관계를 19세기 프랑스 의학을 통해 잠시 살펴보겠습니다. **루이(Louis, P.-C.-A., 1787-18720)**라는 의학자가 주인공입니다. 그는 의학에서 수치적 방법(numerical method)을 퍼뜨린 원조라고 불리는데 특히 19세기 전반기까지 구미 각국에서 널리 이용되던 **사혈법(bloodletting)** 방법의 치료효과를 통계를 가지고 결정적으로 반박한 사람으로 알려져있습니다⁴.

사실 지금도 ‘의술’과 같은 단어가 쓰이고 있습니다만 오랫동안 질병이나 환자에 대한 의사들의 의학적 판단들은 명시적인 규칙으로 나타낼 수 없는 것으로서 일종의 ‘**암묵적인 지식(tacit knowledge)**’로 간주되었다고 합니다. 그런 지식은 의사 개인의 경험에서 나온 것으로서 무엇보다 **환자마다 제각각 다른 개별적 특성을 중요시**하는 관점이었다고 합니다. 그렇게 본다면 많은 사람을 동질적으로 보고 얻은 통계 수치를 가지고 치료법을 비교하고 평가하는 것은 매우 거칠고 위험한 방법이 됩니다. **서로 다른 개인에 대해 아무리 많은 데이터를 얻어보았자 지금 치료해야할 특정 환자의 사례와는 무관하다**는



⁴ Matthews, J. R., *Quantification and the Quest for Medical Certainty*, Princeton UP, 1995.

것입니다. 심지어 라플라스의 후계자로서 19세기 프랑스의 대표적 수학자 중 한사람이었으며 오늘날 ‘포아송 분포’로 널리 알려져있는 Poisson 역시 마찬가지로 의학은 확률을 적용하기에 부적절한 분야라고 생각하였습니다.⁵

그런 적대적인 분위기 속에서 루이는 1830년대에 의학계의 동료들과 논쟁을 벌이면서 통계 수치 데이터를 이용한 치료법의 비교를 통해 임상의들이 더 나은 질병 치료를 할 수 있을 것이라고 주장하였습니다. 물론 그가 통계 데이터만으로 의학연구가 충분하다고 주장한 것은 전혀 아닙니다. 다른 방법들과 더불어 활용한다면 의학이 사실에 바탕을 둔 더욱 과학적인 의학이 될 것이라고 주장한 것이지요.

Poisson은 의학 대신 법원에서 열리는 재판에 배심원으로 참여한 사람들의 판결 데이터를 확률을 이용해서 분석한 바 있습니다. 그는 각 배심원들이 유죄, 무죄 결정을 내리는 것은 앞면 아니면 뒷면이 나오는 동전을 던지는 것과 마찬가지로 보고 보고 유죄 판결에 필요한 배심원 수를 12명 중에서 8명으로 할 것인가, 7명으로 할 것인가라는 문제를 논의했습니다. 그런데 이번에도 배심원들은 다들 사건마다 각각 각자 뚜렷하게 다른 생각을 바탕으로 판단을 하기 때문에 그들의 결정을 단순한 동전던지기처럼 보는 것은 잘못이라는 비판이 나왔습니다. 자세한 내용은 제가 아주 좋아하는 해킹의 책 11장을 보기 바랍니다.⁶

프랑스의 재판 데이터를 분석한 사람은 Poisson 말고도 있었는데 Quetelet도 그중 한 사람이었습니다. 그는 피고인의 유형에 따라 유죄판결율을 비교하였는데 역시 동질적인 집단을 어떻게 결정할 것인가라는 문제와 씨름해야 했습니다. 천문 관측 결과처럼 데이터를 동질적인 집단으로 나누어야 수학자들이 만들어놓은 확률이론을 써서 분석해나갈 수 있었기 때문이지요. 통계 데이터를 가지고 사람들로 이루어진 집단들을 비교하려 했던 19세기 통계전문가들이 고심해야했던 문제는 이처럼 수학 문제가 아니었던 겁니다.

19세기 통계전문가들이 내놓은 집단적 사고의 대표는 아마 케틀레의 ‘**평균인(average man)**’일 것입니다. 평균적인 사람이라는 가상의 존재는 비록 허구이기는 했지만 집단의 여러 특성을 하나의 개체에 구현한 것으로서 매우 호소력 있는 발명품이었습니다.⁷ 하지만 한계도 뚜렷했는데 집단을 평균이라는 단일한 값으로 대표한다는 발상은 동질적인 관측 데이터를 가진 천문학에서 온 것이었기 때문이었습니다. 아무래도 자연과학을 공부하고 천문대에서 일한 케틀레는 사회와 인간을 다소 단순화시켜서 보았던 모양입니다. 사회를 보는 그의 통계학이 더 다듬을 필요가 있다는 것은 여러 사람이 지적했었고 케틀레 자신도 잘 알고 있었던 것 같습니다.

⁵ 해킹, 이언. <우연을 길들이다: 통계는 어떻게 우연을 과학으로 만들었는가?>, 정혜경 옮김, 바다출판사, 2011, 10장.

⁶ 해킹, 이언. <우연을 길들이다: 통계는 어떻게 우연을 과학으로 만들었는가?>, 정혜경 옮김, 바다출판사, 2011, 11장.

⁷ 스티글러, S. M. <통계학의 역사> 조재근 옮김, 한길사, 2005, 제5장.

그런데 사람의 특성들을 측정된 데이터를 가지고 사람을 집단으로 분류하는 작업에 몰두한 사람이 케틀레만은 아니었습니다. 19세기에 엄연한 과학으로 대접받았던 **두개골측정학**이나 **골상학** 등을 생각해보십시오. 두개골 용적으로 인종과 남녀의 지적 수준을 나누고 머리통의 생김새로 사람의 종류를 나누고, 범죄자의 온갖 신체적 특성을 측정해서 범죄자의 유형을 만들어 경찰 수사에 활용하고 아주 재밌습니다. 집단화를 위한 통계데이터의 활용은 어쩌면 19세기를 사로잡은 유행이었던 것 같습니다. 더욱 재미난 것은 천문학에서 온 아이디어로 케틀레가 만든 평균인이나 통계분포라는 개념들(케틀레의 ‘**사회물리학**’)이 19세기 후반 기체의 분자 운동을 연구하는 물리학자들에게 영감을 주었다는 점입니다. 개별 기체 분자가 아니라 집단을 이룬 분자들의 운동을 연구한 J. C. Maxwell이 대표적이라는군요.⁸ 통계적 사고의 역사를 다루는 책에서는 수학보다는 물리학, 생물학, 사회과학 등과 관련된 역사를 만나는 경우가 더 많은 것 같습니다. 개인과 집단의 관계를 하는데 사회학 이야기를 뺄 수 없겠지요.

2.2 자살의 법칙⁹

유럽을 여행한 사람들 가운데에는 영국에 머무는 동안 춥고 음산한 날씨 때문에 고생했다는 사람이 적지 않다고 합니다. 유럽 사람들은 날씨의 영향으로 많은 영국인들이 우울한 성격을 갖게 되며 심지어 자살도 더 많이 할 것이라고 생각했었다는군요. 그런데 오랫동안 상식처럼 받아들여졌던 그런 생각이 19세기에 접어들면서 근거 없는 편견임이 밝혀졌다고 합니다. 영국인들보다는 바다 건너 프랑스인들이 더 자살을 많이 할뿐더러 영국의 자살률은 다른 유럽 나라들과 비교해도 낮은 수준이었기 때문이었지요. 당시 사람들이 자살이라는 문제에 특별한 관심을 쏟았던 이유는 자살을 통해 개인과 사회의 관계를 판단할 수 있다고 생각했기 때문이었습니다. 즉 자살자 수는 각국 사람들의 삶의 질을 비교하고 공동체의 건강 정도를 알려주는 기준으로서 나라의 자존심이 걸린 문제이기도 했겠습니다. 예상과 달리 영국의 자살률이 높지 않다는 사실을 밝힘으로써 영국인의 삶의 질을 다시 평가하게 만든 것은 19세기의 사회통계였습니다.

그런데 영국의 낮은 자살률보다 당시 사람들을 더 놀라게 한 것은 자살이나 범죄를 비롯한 각종 사회통계에서 드러난 규칙성이었습니다. 각국에서 해마다 체계적으로 자살자들을 분류해본 결과

⁸ Porter, T., *The Rise of Statistical Thinking, 1820-1900* (Princeton UP, 1986, Chapter 5.

⁹ 이 부분은 필자가 썼던 다른 글을 수정한 것입니다. “1천만분의 1은 거의 발생하지 않는다?” <나라경제>, 2018. 4. nara.kdi.re.kr.

자살 방법부터가 나라별로 뚜렷이 달랐는데 파리에서는 강물에 투신하거나 질식사하는 방법을 쓴 사람이 일관되게 많았던 반면 런던에서는 그와 전혀 다르게 목을 매거나 총기를 이용해서 목숨을 끊는 사람들이 역시 일관되게 많았다고 합니다. 방법 뿐 아니라 계절별로, 지역별로, 성별로 보아도 자살자 수는 해마다 일정했구요. 지극히 개인적인 결단에 따라 이루어지는 자살이 마치 어떤 정해진 법칙을 엄밀하게 따르는 것처럼 보였던 것입니다.

이런 놀라운 결과로부터 통계를 이용하면 자연현상에서 볼 수 있는 **물리 법칙과 유사한 사회의 법칙**을 찾을 수 있으리라는 생각이 등장했습니다. 만일 사회 현상들이 규칙적인 법칙에 따라 일어난다면 언제 어디서 어떤 일이 얼마나 생길지 미리 예측하는 일도 가능해질 것 같습니다. 그렇다면 혼란이 아니라 질서 있는 사회를 만들 수 있을 것이고 급격한 혁명을 걱정할 필요도 없어질지 모릅니다. 18세기 끝 무렵 일어난 프랑스대혁명과 더불어 산업혁명이라는 큰 변화를 겪던 유럽에서 통계는 사회를 과학적으로 설명하고 예측할 수 있는 유력한 길을 제시하는 것처럼 보였습니다. 각자 독특하고 예측할 수 없는 존재처럼 보이는 개인도 마치 어찌할 수 없는 법칙의 지배를 받으며 사회 속에서 상호작용하고 움직이는 원자 같다고나 할까요?

사회학은 개인이 아니라 집단을 연구합니다. 사회라는 것이 개인들이 모여 만들어지지만 사회 나름의 원리나 법칙에 따라 움직이므로 사회는 단순히 개인의 합이 아니라는 것이지요. 사회학의 역사가 항상 통계데이터와 가깝기만 한 것은 아니었지만 19세기 사회 데이터가 드러낸 규칙성이나 그런 데이터를 해석하는 집단적 사고가 없었다면 사회학은 생기기 어려웠을지도 모릅니다.

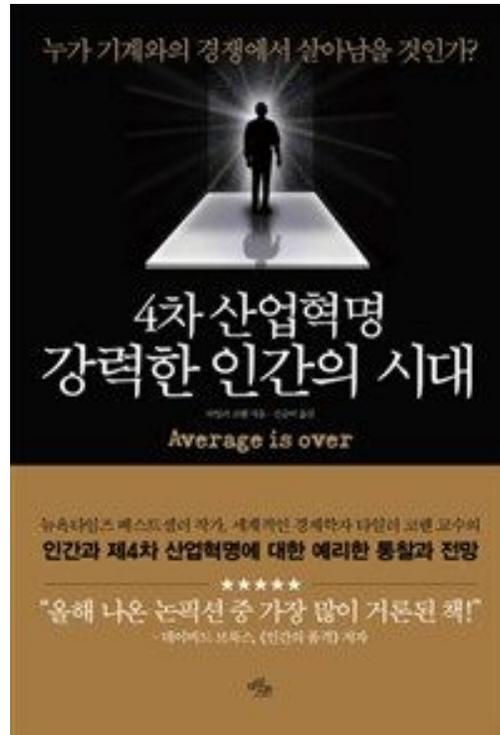
2.3 19세기말 이후의 통계학적 사고: 데이터와 이론의 만남

통계학을 공부한 사람들에게 골턴 이후 통계학의 역사에 등장하는 영국인들의 이름은 낯설지 않을 것입니다. 매우 거칠게 요약해버린다면 19세기 통계의 대세로 자리잡은 집단적 사고를 확률이론과 새로운 분석법들을 이용해서 정교하게 체계적으로 만들어서 비로소 통계학을 과학의 지위에 올린 시기였습니다.

19세기 후반부터 20세기는 통계학적 추론의 시대라고 부를 수 있는 시기였는데 회귀와 상관을 필두로 멋진 개념들이 나타났고 추정, 검정법들이 체계화되고 이전까지의 데이터만으로 접근하기 어려웠던 인과관계에 대한 논의들도 활발해진 시기였습니다. 그렇지만 이 시기의 통계학 역시 개인보다는 집단의 특성을 밝히는 것을 더 중요하게 보았습니다. 매우 많은 진전과 거의 혁명적인 발전에도 불구하고 그런 면에서 20세기는 19세기의 연속이라고 보고 싶습니다.

3. 21세기 빅데이터 시대의 통계학적 사고: 집단에서 다시 개인으로?

4.1 평균에 대한 비판



최근에 나온 평균을 비판하는 책들입니다. 이런 비판은 그리 새롭지 않습니다. 가령 <평균의 종말> 앞부분을 보면 1950년에 미군 전투기 조종사 4천 여명을 대상으로 여러 가지 신체 특성을 측정하여 평균적인 조종사 모형을 만들었다는 이야기가 나옵니다. 문제는 그 모형과 현실 조종사들을 비교해보았더니 모형에 들어맞는 사람이 단 한 사람도 없었다는 겁니다. 따라서 평균적인 조종사의 체격에 맞게 조종석을 설계하면 모든 사람이 불편해진다는 것이지요.

이런 식의 '획일적 평균주의'를 비판한 사례는 드물지 않습니다. 우생학의 창시자이기도 했던 골턴 역시 19세기 후반에 비슷한 예를 들어 케틀레의 평균주의를 비판하고 진부한 평균 대신

특출한 데이터에 주목한 바 있습니다. 우생학의 입장에서는 평범한 사람만 잔뜩 있어서는 사람의 품종개량을 기대하기가 어렵겠지요.

한편 4차 산업혁명 시대라고 불리는 미래의 노동시장에서는 평균에 해당하는 진부한 노동력을 가진 사람의 일자리는 없을 것이라는 주장도 있습니다. 평균이 아니라 특출한 능력이나 특성에 주목해야한다는 비판은 새로운 비판은 아니지만 요즘 유달리 자주 들을 수 있습니다. 우리는 평균에 대한 비판은 개인이 아닌 집단 통계에 대한 비판이라고 읽어도 되겠습니다. 그렇다면 집단이 아닌 개개인이 모두 중요하다 하는 것일까요? 가장 대표적인 사례 중 하나가 개인맞춤형 의학라고 할 수 있습니다. 잠시 앞서 이야기한 19세기 전반기 프랑스 의학을 떠올려 보십시오. 개인의 서로 다른 특성을 강조하는 의학은 200년 전에 죽지 않았던 것입니다.

4.2 개인맞춤형 의학

저는 biostatistics의 최근 연구동향에 대해 사실 1도 모릅니다. 그래도 통계학회 소식지 2018년 4월치에서 생물통계학자인 박태성 교수님의 한국갤럽상 수상 소감을 영어는 빼고 한글로 된 부분만 대충 읽었습니다. 그는 그동안 자신과 연구팀이 이룬 많은 업적과 성과를 자랑한 다음 “최근 화두가 되고 있는 **정밀맞춤의학 구현**에 기여하여 실제 환자들의 삶에 유익한 연구결과를 얻을 수 있도록 연구에 더욱 매진하겠다”는 계획을 밝혔더군요. 그의 뛰어난 연구 업적도 부럽고 훌륭한 포부도 부럽습니당!

그 소감글에서 박태성 교수께서 ‘정밀맞춤의학’이라고 표현한 것은 아마 personalised medicine, precision medicine을 일컫는 것 맞겠지요? Health informatics도 비슷한 의미로 쓰이는 모양이고요. 잠깐 찾아보면 맞춤의학은 종래의 증상에 기반을 둔 질병 진단과 치료의 한계를 극복할 수 있는 새로운 의학이라고 합니다. 무엇보다 집단이 아니라 개개인이 앞으로 걸릴 수 있는 질병을 미리 예측해서 적절한 조치를 할 수 있다는군요. 병에 걸렸을 때 치료 또한 개인의 특성에 맞는 맞춤형 치료가 가능해질 것이라고 하고요. 그런 혁명적인 변화는 방대한 각종 데이터와 그런 데이터를 분석할 수 있는 능력 덕분에 가능하다고 합니다. 그런 능력 가운데 통계학이 매우 중요하다고 합니다.

이 내용은 지금 조금씩 공부하고 있는 중이라서 더욱 부실합니다. 어쨌든 맞춤형의학이라는 말을 들으면 저는 이런 것이 궁금합니다. 질문들만 적어둡니다.

“19세기 이후 통계학이 집단을 단위로 했던 것은 세밀하게 개인까지 분석할 데이터가 없었고 또 적절한 분석방법도 없었기 때문일지 모른다. 그렇다면 집단에 머물지 않고 궁극적으로 개인에까지 닿는 것은 통계학적 분석의 최종 목표일까?”

“만일 의학 뿐 아니라 사회과학에서도 개인까지 분석하게 된다면 집단을 대상으로 분석하고 이론을 만들어왔던 사회학같은 학문의 미래는 없는 것일까?”

“맞춤의학만 하더라도 상당히 융합적인 학문일 수밖에 없을 것이다. 통계학의 미래는 독자적인 존립보다는 섞여서 녹아없어지는 쪽일까? 이미 널리 알려진 데이터과학 등의 다른 이름으로?”

4.3 빅데이터의 시대-집단에서 개인으로, 과연 통계학적 사고의 변화일까?

최근 우리나라 대학 입시에서 통계학과 인기의 인기가 높아졌다고 합니다. 그 이유가 빅데이터의 시대에는 확률, 통계학이 매우 중요하기 때문이라고도 하더군요. 예컨대 빅데이터 분석에는 분류, 군집, 회귀 등 통계학을 공부한 사람에게 익숙한 분석법들이 많이 쓰이고 있습니다. 그러다 보니 초, 중, 고등학교 교육에서 수학 과목의 비중을 줄이고 확률, 통계의 비중을 크게 높여야한다는 주장도 나온다고 합니다.

그런데 그런 주장을 하기는 쉽지만 그게 간단한 문제는 전혀 아니겠지요. 대학에서 지금과 같이 통계학을 전공한 사람들이 통계학개론을 줄여서 중, 고등학생들에게 쉽게 가르치면 될까요? 지금 수학처럼 수능에 많이 나오니 무조건 확률, 통계를 공부하라고 가르쳐선 아닐 것 같습니다. 수포자에 더하여 통포자까지 우수수 쏟아질 수도 있겠습니다. 어쩌면 ‘통계학적 사고’라는 주제를 통계학회에서 더 자주 논의해야 할 이유가 여기에도 있을 것 같습니다. 지금까지는 수학교육학회 등에서 그런 주제에 대해 더 많은 논의를 해왔던 것 같습니다.

앞에서 저는 21세기 통계학적 사고의 변화를 “집단에서 개인으로”라고 간단히 표현한 바 있습니다. 저는 빅데이터 분석에서 많은 통계학적 분석법이 활용된다는 사실과, 보다 근본적인 바탕을 흐르는 사고의 변화 사이에 어떤 관계가 있을까 궁금하였습니다.

빅데이터분석의 방향이 각 개인의 특성까지 섬세하게 헤아리는 방향이고 거기에 각종 (오래된, 또는 새로 개발된) 통계학적 분석들이 많이 쓰인다고 해서 그것을 과연 통계학적 사고의 변화라고 보아도 될지요? 그것이 과연 통계학적 사고의 변화이거나 한 것일까요? 통계학은 과연

최근 빅데이터 붐의 핵심에 들어 있거나 한 것일까요?

그래도 통계가 각광을 받았던 19세기에 통계는 개인을 살펴서는 알 수 없는 집단의 특성을 파악한다는 근본적인 사고의 차원에서 중요한 방법으로 뚜렷하게 변화의 핵심에 있었습니다. 그런데 오늘날 빅데이터의 시대에 통계학은 혹시 사고의 차원까지 올라가지 못한 채 전문적인 도구의 역할을 하는 데 그치고 있지는 않을까요? 도구의 역할을 하는 통계학이라면 쓸모있는지 여부만 중요할 뿐 굳이 그 학문의 바탕을 이루는 이런저런 통계학적 사고 같은 것을 생각할 필요도 없겠고 frequentist-Bayesian 논쟁 같은 것도 더 이상 필요 없겠지요. 이 모두가 학문의 세분화, 전문화 시대에는 어쩔 수 없는 일일까요?

제가 앞에서 잠깐 언급한 책 <국가처럼 보기>에는 “왜 국가는 계획에 실패하는가?”라는 부제가 붙어 있습니다. Seeing Like a State라는 제목에서 Seeing Like a Statistics라는 표현이 떠올랐다는 것은 국가의 집단적 사고가 실패하였듯 지난 19세기 이후 집단을 중심으로 세상을 보는 통계학적 사고방식의 유효기간도 어쩌면 그리 길지 않을지 모르겠다는 생각이 들었기 때문입니다. 요행히 통계학은 빅데이터 덕분에 개체의 고유한 특성까지 헤아리는 세밀한 분석법을 활용해서 과거 국가들이 겪은 것과 비슷한 실패를 잘 피할 수 있을까요?