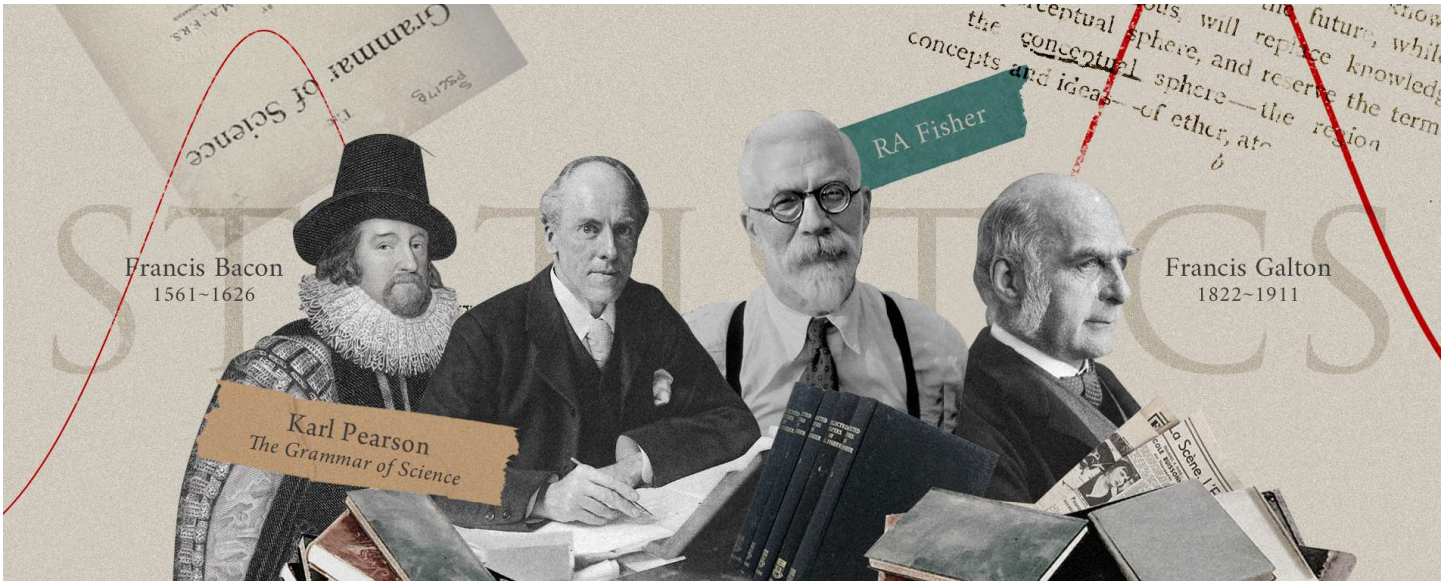
통계학

통계학 이야기

2020년 9월 7일

김재광



들어가며

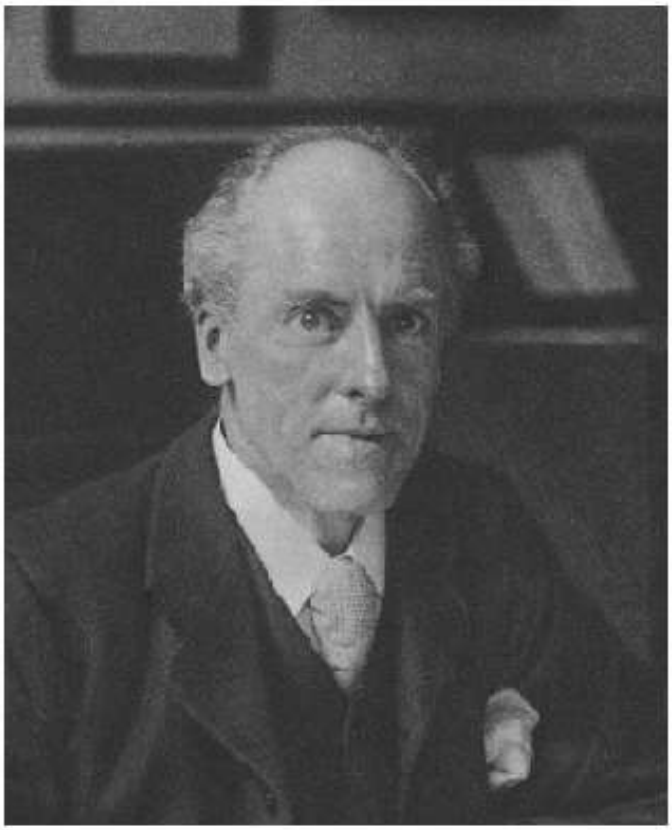
세상이 복잡해질수록 불확실성은 높아지는데 이러한 불확실성 가운데에서도 올바른 판단을 하기 위해서는 확률에 기반한 사고가 유용합니다. 이를 좀 더 시적으로 표현하면 "신이 내린 불확실성의 저주를 푸는 열쇠가 바로 확률"이라고 할 수 있을 것입니다. 불확실한 미래는 다가올 미래에 대한 확률을 계산함으로써 최선의 대응을 할 수 있는 것이고, 불확실한 정보는 그 정보가 어떤 현상에서 나왔을지에 대한 확률을 계산함으로써 최선의 판단을 할 수 있는 것입니다. 이러한 확률은 홀로 존재하는 것이 아니라 데이터와 만나서 구체적으로 계산되어야 유용해질 텐데 이에 대한 과학적 방법을 연구하는 학문이 통계학입니다. 확률 자체는 수학의 영역이지만 데이터와 확률의 만남은 통계학의 영역인데, 이는 수학과는 제법 다른 자세를 요구하기에 통계학은 수학과 분리되어 하나의 독립된 학문으로 활동하기 시작했습니다. 세계 최초의 통계학과가 1911년에 영국의 유니버시티 칼리지 런던University College London에서 처음 만들어졌으니 아직 젊은 학문이라고 볼 수 있고 앞으로도 발전가능성이 높은 학문이라고 할 수 있을 것입니다.

통계학의 역사는 상대적으로 짧지만 데이터가 거의 모든 분야에서 발생하고 있고 데이터를 이용하여 보다 나은 의사결정을 하고자 하는 욕구는 보편적인 것이기에, 통계학은 지난 100년 동안 급속도로 전파되고 활발하게 사용되었습니다. 이제는 통계학이 너무 광범위하게 퍼져서 각 도메인에서 어떻게 활용되고 다른 분야들과 어떠한 상호작용을 거치면서 발전해 왔는지를 한 사람이 조감하는 것은 매우 어려운 일입니다. 하지만 이러한 어려움에도 불구하고 조심스럽게 통계학에 대한 큰 그림을 보여주려는 시도는 충분히 가치있는 일일 뿐만 아니라 시의적절한 일이라고 생각합니다.

이를 위하여 저는 먼저 통계학이 갖는 철학적 의미를 서술하고, 그다음 과학으로서 통계학의 발전을 이야기 한 후에, 앞으로 기대되는 공학으로서 통계학의 미래를 말씀드리고자 합니다. 이러한 내용들이 제한된 지면에도 불구하고 많은 분들에게 통계학에 대한 이해와 흥미를 높일 수 있는 계기가 되기를 희망합니다.

철학으로서의 통계학

통계학이 인류의 지성사에서 어떠한 의미를 갖고 어떠한 철학적 좌표에 위치하고 있는가는 저의 오래된 질문이기도 했는데, 저는 해답의 실마리를 "왜 통계학이 영국에서 시작되었는가?"에서 찾을 수 있다고 생각합니다. 통계학은 지금으로부터 약 100여 년 전에 칼 피어슨Karl Pearson,1857-1936과 로널드 피셔RA Fisher,1890-1962라는 두 영국인에 의해 정립되기 시작한 도구 학문입니다. 칼 피어슨Karl Pearson은 세계 최초로 통계학과를 설립하신 분이기도 한데, 그의 스승은 골턴Francis Galton, 1822-1911경으로 상관계수와 측정의 개념을 확립한 분이기도 하고『종의 기원』을 쓴 찰스 다윈의 사촌이기도 합니다. 이렇게 학문적 뿌리를 거슬러 올라가면 결국 만나게 되는 사람은 경험주의 철학자 프란시스 베이컨입니다. 경험주의empiricism는 감각의 경험을 통해 얻은 증거들로부터 비롯된 지식을 강조하는 이론입니다. 감각 경험이란 결국 현실 데이터를 지칭합니다.




좌 그림1 칼 피어슨(1910년)

우 그림2 로널드 피셔(1913년)

wikimedia(그림1,그림2)

철학사에서 중세를 벗어나 근대를 여는 데는 베이컨과 데카르트의 활약이 큽니다. 베이컨은 경험주의를 창시했고 이는 귀납적 사고를 중시합니다. 반면 데카르트는 합리주의를 창시했고 이는 연역적 사고를 중시합니다. 이전 중세에는

권위와 전통에 기대어 지식을 알려고 했다고 하면 근대에서는 권위와 전통을 부정하고 오로지 경험과 이성을 통해서 새롭게 지식 체계를 세우고자 합니다. 종교의 영향력에서 벗어나 기존의 지식을 의심하고 참된 지식을 세우려는 노력에서 필요한 도구는 결국 경험(데이터)와 이성(논리)라는 것입니다. 베이컨에서 시작된 경험주의는 영국에서 계속 발전하여 존 로크, 조지 버클리, 데이비드 흄으로 계승됩니다. 그들은 모든 올바른 지식은 감각적 경험으로 유래된다는 견해를 가졌는데요, 이러한 경험주의의 단점은 자칫하면 회의주의에 빠진다는 것입니다. 왜냐하면 지식이 데이터로부터 얻어진다는 것은 인정하지만 우리가 모든 데이터를 다 관측할 수는 없기에 그 지식을 결코 확신할 수 없다는 난관에 부딪히고, 따라서 확신하지 못하는 지식을 얼마나 신뢰할 수 있겠느냐는 문제가 생깁니다.

이런 문제를 수리적인 방법으로 해결한 것이 결국 통계학입니다. 통계학은 제한된 관측으로 얻어지는 결론에 대한 확실성의 정도를 확률이라는 개념으로 설명해 냅니다. 그리고 확률이라는 것을 계산하는 과정에서 수리적 논리를 부여하여 보편성을 획득합니다. 따라서 경험주의자 입장에서는 100% 확신하는 지식은 없지만 통계학을 통해서 어떤 지식이 상대적으로 더 확실한 것인지에 대해 합리적인 진술을 할 수 있게 됩니다. 또한 통계적 가설검정은 일종의 귀류법으로 볼 수 있습니다. 통계학에서 어떤 가설을 데이터로 증명하기보다는 그 가설을 일단 부정한 후에 (이를 귀무가설이라고 합니다) 데이터가 그 귀무가설을 확률적으로 강하게 반증하는 경우 처음 가설을 채택하는 논리를 따릅니다. 그래서 통계학은 경험주의의 한계를 극복하고자 하는 인간 지성사의 노력의 산물입니다. 이러한 통계적 추론의 이론적 틀을 마련한 분이 로널드 피셔RA Fisher이니 그분은 인류에게 큰 선물을 한 것입니다. 마치 뉴턴이라는 천재가 물리학의 기초를 닦아서 인류에게 큰 선물을 한 것처럼 피셔Fisher 선생 역시 통계학의 기초를 닦아서 인류에게 큰 선물을 한 것입니다.

과학으로서의 통계학

이러한 통계학이 과학으로서의 위상을 갖기 위해서는 통계학이라는 현실 속의 문제를 실질적으로 해결하기 위해 진화할 것이 요구됩니다. 실제로 20세기는 통계학이 과학으로서의 위상을 확립한 시기라고 볼 수 있을 것입니다. 통계학적 방법론들이 각 도메인에 적용되어 경험적 지식들이 늘어나고 체계를 갖추게 되었습니다. 가장 대표적인 것은 농업과 의학에서의 성공이라고 할 수 있습니다. 실험계획법이라는 통계학 방법을 통해 농업기술과 비료 등이 개발되고 신약 개발이 과학적으로 검증되어 수많은 질병과의 전쟁에서 승리할 수 있었습니다. 또한 통계학은 경영과 행정에도 혁신을 가져왔습니다. 데이터를 기반으로 한 품질관리를 통해서 제조업의 혁신이 이루어졌고 확률표본 추출을 통한 샘플링을 통해 근거기반 행정이 자리 잡을 수 있게 되었습니다. 그 외에도 교육학, 심리학, 경제학, 사회학, 정치학 등에 통계학이 끼친 영향은 이루 말할 수 없이 큽니다. 통계학적 방법론을 과학에 도입한 칼 피어슨Karl Pearson의 유명한 저서의 제목은『The Grammar of Science』이기도 했습니다. 통계학은 과학의 문법이라는 것입니다.

통계학이 과학의 방법론이긴 하지만 통계학을 사용했다고 해서 결론이 다 타당하거나 과학적인 것은 아닙니다. 통계학이 과학으로 자리잡기 위해서는 두 가지 중요한 전제가 성립해야 합니다. 첫 번째로 통계학적 지식 체계가 논리적으로 오류가 없어야 합니다. 통계학적 지식체계라고 함은 통계학에 사용되는 분석 방법론들이 다루고 있는 가정에서 결론으로까지의 여정을 의미합니다. “만약 A라는 조건을 만족하면 B라는 성질을 만족한다.”로 표현되는 통계학적 지식체계는 통계학의 주요 커리큘럼의 내용이기도 합니다. 이러한 전문 지식은 주로 수학적 논리를 사용하여 연구됩니다. 두 번째 중요한 전제는 통계학적 지식에서 사용되는 가정들이 현실에 부합해야 합니다. 비현실적인 가정을 바탕으로 논리적으로 완벽한 방법으로 얻어진 결론은 그 결론이 아무리 유용하다 할지라도 쓸모가 없습니다. 여기에서 통계학이 순수수학과는 다른 가치관을 가지고 있음을 알 수 있습니다. 순수수학은 그 자체의 아름다움을 추구한다고 한다면

통계학은 그보다는 유용성과 적용 가능성에 더 가치를 둡니다. 그렇다면 통계학의 가정들이 현실에 부합해야 한다는 말이 구체적으로 의미하는 것은 무엇일까요?

"

순수수학은 그 자체의 아름다움을 추구한다고 한다면

통계학은 그보다는 유용성과 적용 가능성에 더 가치를 둡니다.

"

통계학에 사용되는 가정은 크게 두 가지로 나뉩니다. 하나는 데이터에 대한 가정이고 다른 하나는 모델에 대한 가정입니다. 논리학에서는 어떤 결론을 얻어내기 위해서 논거, 전제, 논리가 결합이 되어야 하는데 이를 통계학적으로 표현하자면 논거는 데이터이고, 전제는 모델이며, 논리는 통계 방법론이 되는 것입니다. 이 삼박자가 잘 맞아야 올바른 결론을 얻게 되는 것입니다. 이 중에서 제일 중요한 것을 하나 고르라면 아마도 데이터가 아닐까 생각합니다. 아무리 훌륭한 모델과 좋은 방법론을 가지고 있더라고 데이터가 쓰레기이면 결국 쓰레기와 같은 결과가 얻어지는 것입니다. 이는 아무리 훌륭한 요리사와 좋은 요리법이 있다고 하더라도 식재료가 엉망이면 좋은 음식이 나올 수 없는 것과 마찬가지입니다.

따라서 과학으로서의 통계학이 전제하고 있는 첫 번째 요인은 질 좋은 데이터입니다. 그렇다면 질 좋은 데이터라는 것은 무엇을 의미할까요? 그걸 위해서는 우리가 원하는 결론이 포함하고자 하는 타당성의 범위를 이해해야 합니다. 우리가 원하는 결론이 지칭하고자 하는 집단을 자료가 충분히 대표하고 있는지에 대한 판단이 필요할 뿐만 아니라 얻어진 측정 방식이 원하는 개념을 정확하게 반영하고 있는지에 대한 판단도 필요합니다. 즉, 자료의 대표성과 측정의 정확성이 양질의 데이터를 결정짓는 중요한 두 가지 요인입니다. 측정의 어려움은 특히 사람을 대상으로 자료를 얻는 경우에 더욱 흔히 발생합니다. 이는 인간을 대상으로 실험하는 경우 윤리적인 이유로 실험이 어려울 뿐만 아니라 설문 등을 통해 측정을 하는 경우 간섭이 일어나서 측정오차가 생긴다는 것입니다. 그래서 이러한 문제를 해결하기 위해 데이터 자체를 얻어내는 과정 자체를 과학적으로 설계할 필요가 있는데 이는 자료 수집 비용의 증가를 의미하기도 합니다. 그런 양질의 데이터를 얻기가 힘들다면 적어도 사용하고자 하는 데이터의 증거 적절성 여부를 우선 판단해야 하는데 초보자들은 이에 대해 어려움을 겪거나 아예 그 문제 자체를 인지하지 못합니다. 데이터의 적절성을 판단하는 것은 법정에서 증거가 제시되었을 때 그것을 증거로 채택할 것인지 아닌지를 먼저 판단하는 것과 비슷한 것입니다.

두 번째 전제는 모델의 적절성입니다. 첫 번째 전제를 통과한 데이터를 사용하여 확률을 계산하기 위해서는 사용하고자 하는 확률 모형의 적절성 또한 전제되어야 합니다. 모형은 현실에 대한 이해의 틀이고 모형의 적절성은 결국 그 모형이 데이터의 현실을 얼마나 잘 반영하느냐에 따라 판단될 수 있는데, 이는 자료의 구조를 이해해야 함을 의미합니다. 자료가 범주형인지 연속형인지, 시계열 자료인지, 공간자료인지, 계층적 구조를 가진 것인지 아닌지 이러한 정보들은 자료의 구조에 대한 이해를 전제로 하고 이것은 모형의 적절성을 결정합니다. 현실(데이터의 구조)을 제대로 반영하는 모형을 사용할수록 그로부터 얻어지는 결론의 정확성은 높아질 것입니다. 이러한 두 가지 전제를 만족하여 얻어진 통계학적 결론은 상당한 수준의 과학적 근거를 가지게 됩니다.

공학으로서의 통계학

통계학이 하나의 학문으로서 진화의 운명에 처해있다는 것은 21세기에 외부 환경의 변화에 적응해야 하는 숙제가 있음을 의미하기도 합니다. 컴퓨터와 인터넷의 보급으로 인해 통계학 역시 위기와 기회요인을 맞이하고 있는데요, 먼저 위기요인은 전산학의 머신 러닝Machine Learning과의 경쟁을 피할 수 없다는 것입니다. 20세기에 크게 발전한 과학으로서의 통계학은 과학적 엄밀성의 측면에서는 공학적 접근법에 우위를 차지하고 있지만 데이터를 통해 지식을 발견하는 과정의 주체가 인간임을 전제로 하는 것이기에 전문지식을 가진 인간의 끊임없는 개입과 해석을 요구하고 있습니다. 반면 전산학에서는 학습의 주체가 사람이 아닌 컴퓨터(기계)이기에 인간의 개입을 최소화하고 있습니다. 따라서 통계학은 과학적 우위를 점령하고 있는 대신, 자동화를 통한 대량생산을 추구하는 공학적 측면에서는 열위에 놓여 있는 것입니다.

> 기계 학습에서 데이터의 편견 문제는 사실 통계학의 대표성이나 측정 문제와 밀접한 연관이 있고,
>
> 따라서 데이터의 편견을 보정할 때 인간의 개입이 필요할 경우 통계학적인 접근법이 사용될 수 있습니다.

하지만 이러한 변화 가운데에서 저는 통계학의 기회 요인 역시 생기고 있다고 생각합니다. 먼저 머신러닝 방법론이 확장되는 과정에 통계학적 시각이나 방법론들이 기여할 부분이 많이 있을 것입니다. 또한 기계 학습에서 데이터의 편견 문제는 사실 통계학의 대표성이나 측정 문제와 밀접한 연관이 있고, 따라서 데이터 편견을 보정할 때 인간의 개입이 필요할 경우 통계학적인 접근법이 사용될 수 있습니다. 또한 도메인 지식과 데이터 기반 지식을 결합하는 문제 역시 일종의 통계학적 관점으로 이해될 수 있습니다. 즉, 기계학습이나 인공지능을 컴퓨터가 스스로 학습하는 것으로 국한하는 것이 아니라 인간이 인공지능을 탑재하여 강화된 판단의 도구로 사용하는 개념으로 보게 될 때 다시 인간 중심의 통계학이 부활할 수 있는 것입니다. 결국 두 학문 분야가 서로 경쟁한다기 보다는 서로 경계가 허물어지고 융합되고 있다는 것이 더 정확할 것입니다. 아직은 21세기의 초반이니 공학으로서의 통계학의 발전은 초기 단계에 있다고 할 수 있을 것입니다.

마치며

저는 통계학이라는 학문이 어떠한 뿌리를 가지고 있고 어떠한 과정을 거쳐서 앞으로 발전을 하게 될지에 대해 다소 주관적인 감상을 정리해 보았습니다. 통계학이라는 것이 현실과 유리된 것이 아니고 데이터를 통해서 얻어지는 수많은 현실 문제들을 어떻게 바라보고 어떻게 해결할 수 있을지에 대한 접근법을 제공한다는 측면에서 현대인의 필수 교양이라고 생각합니다. 또한 통계학은 홀로 서 있는 학문이 아닌 다른 도메인 학문을 도와주는 도구학문이므로 도메인 전문가와 통계 전문가의 협업을 통해 발전하는 것이므로 협업이 장려되는 문화에서 더욱 화려하게 꽃을 피울 수 있을 것입니다.

저는 또한 데이터라는 것이 측정을 바탕으로 숫자화된 것이고 그 측정이라는 것이 해당 개념을 전제로 하는 것임을 강조하고 싶습니다. 우리가 이미지를 데이터화할 수 있는 것은, 이미지를 화소로 분해해서 숫자로 표현한 뒤 이를 이진법으로 바꾸어 데이터로 저장할 수 있기 때문입니다. 이러한 과정에 이미 광학적 지식이 반영된 것입니다. 데이터로 변환되지 않는 것은 컴퓨터에서 학습되지 못하는 것이고 그것으로부터 어떤 예측이나 지식을 쌓지 못하게 됩니다. 그래서 측정은 해당 속성에 대한 것을 숫자로 그 강도를 표현할 수 있어야 하는데 이것이 가능하려면 해당 개념을 바탕으로 얻어지는 것입니다. 후각이 시각보다 측정이 더 어려운 것도 후각자료에 대한 이해도가 시각자료에 대한 이해도보다 더 어렵기 때문입니다.

이렇게 개념이라는 것은 인지적 성숙도의 문제이기도 합니다. 우리가 보라색이라는 개념이 있어야 보라색을 인지할 수 있는 것처럼 많은 사회적 과제들이 사실은 개념의 세분화를 바탕으로 인식되고 그에 대한 조작적 정의와 측정의 타당성을 논의하는 것이 필요한데 그러한 문제들이 통계 전문가들의 힘만으로 해결될 수 있지 않습니다. 주어진 데이터를 분석하는 것은 머신러닝이나 통계학으로 해결될 수 있을지 모르겠지만, 필요한 문제를 해결하기 위해 문제를 어떻게 정의하고 데이터를 어떻게 수집할 것인가는 지식산업의 성숙도에 따라 결정되는 것입니다.

따라서 지식산업은 이러한 지식 생태계가 성숙한 곳에서 잘 발달할 수 있을 텐데, 이러한 지식 생태계는 관련 인력이 풍부하고 수준이 높아야 할 뿐만 아니라 커뮤니케이션의 비용이 낮아야 하는데 이를 위해서는 상호 신뢰와 존중이 바탕이 되어야 합니다. 즉, 지식생태계를 성숙시키기 위해서는 그에 걸맞는 문화가 갖추어져야 한다는 것입니다. 대학이 이러한 문화 성숙에 큰 역할을 해야 하는 것은 당연할 것입니다. 아무쪼록 이러한 지식 생태계가 한국에서도 잘 만들어지게 되길 기원합니다.