7장

관계의 이해

상관관계와 인과관계

아이스크림을 많이 먹으면 걸리는 병

지금은 잊혔지만 20세기 중반까지 인류를 꾸준히 괴롭힌 질병이 있었습니다. 바로 소아마비입니다. 소아마비는 어린아이가 특정한 바이러스에 감염되어 근육이 마비되는 병이며, 소아마비에 걸리면 잘 걷지 못하게 됩니다. 인류는 소아마비의 원인을 찾기 위해서 노력했고, 20세기 중반에 백신이 개발되어 지금은 거의 사라졌습니다. 백신을 개발할 당시 많은 데이터를 분석한 결과 연구진은 흥미로운 사실을 발견했습니다. 아이스크림을 먹는 아이가 많아질수록 소아마비에 걸리는 아이도 많아진다는 것이었습니다. 아이스크림 판매량이 높아지면서 소아마비 환자의 수가 증가하는 것을 데이터를 통해 발견한 것입니다. 아이스크림 성분에 소아마비를 유발하는 유해물질이 들어 있는 것이 틀림없어 보였습니다.

그러나 아이스크림과 소아마비의 관계는 해프닝으로 밝혀졌습니다. 사실 문제는 수영장이었고, 수영장에서 물을 통해서 소아마비 바이러스가 쉽게 전염되었던 것입니다. 수영장은 주로 여름에 가고 아이스크림 또한 여름에 많이 먹으니, 여름이 매개체였습니다. 여름에 소아마비 환자도 증가하고 아이스크림 판매량도 증가합니다. 청량음료 판매량과 소아마비 환자의 수도 상당히 밀접한 관계가 있습니다. 둘 다 여름에 늘어나기 때문입니다.

상관관계와 인과관계

데이터를 통해 얻은 관계를 상관관계라고 합니다. 우리가 알고 싶어 하는 것은 인과관계입니다. 소아마비 사례는 상관관계에서 인과관계를 확인하는 것이 만만치 않은 문제라는 것을 잘 보여줍니다. '손을 씻지 않으면 병에 걸린다'는 지식은 지금은 아무도 의문을 품지 않는 상식입니다. 하지만 이 상식은 19세기 중반에야 발견되었으며, 이 단순한 상식은 다른 어떤 의학적 발전보다 인류의 수명을 가장 많이 늘린 발견으로 평가받습니다.

손 씻기의 중요성을 설파한 사람은 헝가리 출신 의사 이그나즈 제멜바이스Ignaz Philipp Semmelweis였습니다. 19세기 중반까지도 위생이라는 개념이 없어서 병원의 의사들이 손뿐 아니라 수술도구도 닦지 않고 재사용했습니다. 제멜바이스는 병원에서 분만한 산모의 사망률이 산파를 통해서 분만한 산모의 사망률보다 3배나 높은 것을 발견합니다. 이러한 데이터 분석을 바탕으로 제멜바이스는 병원과 산파의 차이를 확인한 뒤 손 씻기의 중요성을 설파합니다. 하지만 당시 의사들은 '손을 씻지 않는 의사는 살인자'라고 비판하는 제멜바이스를 좋아하지 않았습니다. 의사들은 자신들의 약점을 들추는 제멜바이스의 주장을 엉터리라고 비난하며 무시했습니다. 이에 상처받은 제멜바이스는 결국 정신병원에서 생을 마감합니다. 이후 위생에 대한 데이터 추적과 다양한 연구를 통해서 손 씻기와 건강은 인과관계로 받아들여졌습니다. 손 씻기라는 너무 당연한 상식이 인과관계로 받아들여지는 것도 순탄하지만은 않았던 것입니다.

이처럼 인과관계를 밝히는 것은 생각보다 매우 어렵습니다. 데이터를 통해서 우리가 알 수 있는 것은 대부분 상관관계입니다. 그리고 많은 상관관계 중 상식으로는 인과관계처럼 보이지만 실제는 아닌 예를 우리 주위에서 쉽게 찾아볼 수 있습니다. 가구당 자동차 보유 대수와 차량당 주행거리의 관계는 흥미롭습니다. 언뜻 생각하기에 자동차를 많이 보유한 가정은 차량을 나눠 타기 때문에 각 차량당 주행거리가 감소할 것이라고 생각할 수 있습니다. 그러나 데이터 분석에 의하면 보유 대수가 늘어나면 주행거리도 늘어납니다. 이 상관관계를 인과관계로 설명하는 이론 중 하나로 자동차를 많이 보유하면 운전하고 싶은 욕망이 늘어나서 주행거리가 늘어난다는 해석을 합니다. 하지만 좀 더 합리적인 설명은 자동차 보유 대수가 주행거리 증가의 원인이 아니라, 주행거리가 긴 가정에서 자동차를 많이 구입한다는 것입니다. 즉, 자동차 보유 대수가 차량당 주행거리의 원인이 아니라 주행거리가 보유 대수의 원인이 되는 반대의 인과관계인 것입니다. 어느 주장이 더 합리적인지는 추가 연구가 필요하지만 상관관계와 인과관계는 이처럼 오묘한 면이 많습니다.[8]

또 다른 사례로는 이혼과 수명의 관계입니다. 국내 한 연구에 의하면 이혼한 사람의 수명이 이혼하지 않은 사람의 수명에 비해서 무려 8~10년

더 일찍 죽는다는 것이 밝혀졌습니다. 그리고 이 결과를 바탕으로 '이혼이 주는 정신적 충격이 건강에 해가 된다'라는 지극히 상식적인 인과관계 설명이 가능해 보입니다. 하지만 이 주장에는 많은 허점이 있습니다. 이혼 자체가 아니라 이혼하는 이유가 건강과 관계될 수 있기 때문입니다. 예를 들면 경제적으로 어려운 사람이 이혼율이 높고 건강도 안 좋을 수 있습니다.

표 1 버클리대학교 입학 데이터

	합격	불합격	지원자(계)
남자	1400(52%)	1291(48%)	2691(100%)
여자	772(42%)	1063(58%)	1835(100%)
전체	2172(48%)	2354(52%)	4526(100%)

표 2 단과대별 합격률(회색은 여학생 합격률이 높은 단과대)

	남		여	
분야	지원자(명)	합격률(%)	지원자(명)	합격률(%)
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	60	341	70

1951년에 미국의 통계학자 에드워드 심슨Edward H. Simpson이 재미있는 논문을 발표합니다. 데이터를 분석하는 방법에 따라서 상관관계가 정반대로 나온다는 것입니다. 〈표 1〉은 버클리대학교 입학 사정 결과입니다. 남학생이 여학생에 비해서 합격률이 더 높습니다. 이 표만 보면 입학에 성차별이 존재하는 것 같습니다. 그런데 같은 데이터를 지원한 전공분야(예: 법학, 의학, 공학, 과학 등)별로 정리하면 결과가 완전히 바뀝니다. 〈표 2〉의 결과를 보면 전공별 합격률은 여학생이 더 높습니다. 6개의 단과대학 중 4개의 전공에서 여학생의 합격률이 더 높았습니다. 이렇게 결과가 뒤바뀐 이유는 여학생은 합격이 어려운 전공에 많이 지원했기 때문입니다(E분야). 성별이 합격률에 영향을 미치지만, 단순히 버클리대학교가 남학생을 선호하는 것은 아닙니다. 오히려 왜 여학생이 합격하기 어려운 전공에 많이 지원했는지에 대한 진지한 고민이 필요할 것입니다. 이렇듯 데이터로 인과관계를 알아내는 것은 굉장히 어렵고, 때론 거의 불가능해 보이기까지 합니다.[9]

인과관계 알아내기: 실험

—

데이터를 통해 인과관계를 알 수도 있습니다. 바로 '임의화 실험'Randomized experiment이라는 특별한 방법을 이용하여 데이터를 모으면 됩니다. 담배와 폐암의 관계를 이용하여 설명해보겠습니다. 1990년대 후반부터 미국에서 흡연에 대한 경각심이 매우 커졌습니다. 제가 처음으로 미국 여행을 갔던 1993년에 비행기 좌석에서 흡연이 가능했던 것을 기억해보면, 담배가 얼마나 빨리 우리 사회에서 퇴출되고 있는지 느낄 수 있습니다. 1960년대에는 미국에서 임산부의 70퍼센트 이상이 흡연을 했다는 놀라운 통계도 있습니다. 현재 담배가 폐암의 원인이라는 것은 대부분의 선진국에서 받아들여지고 있는 듯합니다. 하지만 담배가 폐암의 원인이라는 것은 과학적으로 아직 완전히 증명되지 않았습니다. 흡연으로 인한 폐암

환자들이 담배 회사를 상대로 소송한 재판에서 승소하는 경우가 매우 드문 것도 이 때문입니다. 국내에서는 건강보험공단이 담배 회사에 500억 원대의 손해배상청구 소송을 걸었다가 2020년 11월에 패소하기도 했습니다.

데이터를 분석해보면 담배를 피우는 사람이 폐암에 걸릴 확률은 매우 높습니다. 이 결과만 보면 담배가 폐암의 원인이라는 인과관계에는 이견이 없어 보입니다. 그러나 담배를 피우는 사람이 경제적・사회적으로 약자인 경우가 많고, 열악한 환경에서 일할 확률이 높습니다. 따라서 담배가 폐암의 원인이 아니라 담배를 피우는 사람의 경제적・사회적 열악한 환경이 폐암의 원인일 수 있습니다. 실제로 담배 회사는 이 논리를 법정에서 주장하는데, 이를 반박하기 쉽지 않습니다.

담배가 폐암의 원인이 되는지 직접적으로 알 수 있는 방법은 무작위로 선택된 사람들에게 담배를 피우도록 하는 실험을 통해서 데이터를 모으면 됩니다. 무작위로 선택되었기 때문에 경제적・사회적 위치가 한쪽으로 치우치지 않으며, 이렇게 모은 데이터에서 흡연자가 폐암이 걸릴 확률이 높게 나오면 담배가 폐암의 원인이라고 판단할 수 있습니다. 겨울에도 아이스크림을 먹어보면 아이스크림이 소아마비의 원인인지 알 수 있는 것과 같은 이치입니다. 물론 이러한 임의화 실험을 통한 데이터 수집은 인권 침해 등의 이유로 사람에게 실제 적용이 불가능합니다. 대신 사람이 아닌 동물에 대해서는 가능하고, 널리 사용되고 있습니다. 제약 회사에서 신약을 동물에 적용하여 안정성 및 효율성에 대한 데이터를 모을 때 임의화 실험을 사용합니다. 공장에서 최적의 공정 조건을 찾을 때도 임의화 실험을 통해서 데이터를 수집합니다.

실험 데이터의 수집을 위한 매우 다양한 임의화 방법이 존재하고 실제 실험에서 최적의 임의화 방법을 선택하는 것은 전문가만이 할 수 있습니다. 자동차 연비를 비교하는 실험을 구상해봅시다. H사의 M모형과 S사의 T모형의 연비를 비교하려고 합니다. 각 모형별로 5대를 실험에 사용하여 서울에서 부산까지 주행하고 연비를 측정합니다. 실험에 참여하는 운전자 5명이 있고 각 운전자는 2번씩 운전할 예정입니다. 운전자를 자동차에 어떻게 배정하는가가 이 실험의 핵심입니다. 총 10번의 주행에 임의로 운전자를 2번씩 배정할 수 있습니다. 또는 각 운전자에게 M모형 1대와 T모형 1대를 무작위로 배정할 수 있습니다. 이 2가지 임의화 방법은 각각 장단점이 있습니다.

이렇게 다양한 임의화 방법의 개발 및 장단점에 대한 연구를 하는 분야를 '실험계획법'이라고 하는데, 데이터과학에서 매우 중요한 주제입니다. 실험 데이터를 많이 모으는 제조업이나 제약업 등에서 널리 사용됩니다. 특히 제조업에서 최상의 품질을 위한 최적의 공정 조건을 찾는 문제와 제약업에서 신약의 효과를 증명하기 위한 임상시험에서 유용하게 적용됩니다.

정리하자면 데이터에는 2가지 종류가 있습니다. 일어나고 있는 현상을 관찰하여 정리한 데이터와 임의화 실험을 통해서 얻은 데이터입니다. 전자를 관측 데이터, 후자를 실험 데이터라고 합니다. 예를 들어 인구조사는 관측 데이터이고, 동물시험으로 얻는 데이터는 실험 데이터입니다. 관측 데이터에서는 상관관계만 알 수 있는 반면에 실험 데이터에서는 인과관계를 알 수 있습니다. 그런데 실험 데이터는 모으기 어렵습니다. 비용도 많이 들고 윤리적인 문제도 해결해야 합니다. 특히 사람을 대상으로는 거의 불가능합니다. 반면에 관측 데이터는 다양하게 모을 수 있습니다. SNS에서 얻는 데이터도 관측 데이터입니다. 공장에서 모으는 센서 데이터도 관측 데이터입니다. 금융회사에서 보유하는 신용 관련 데이터도 관측 데이터입니다. 관측 데이터로부터 인과관계를 파악할 수 있는 데이터 분석 방법론이 지금도 계속 개발되고 있습니다. 특히 여기에는 빅데이터가 큰 역할을 합니다.

인과관계 살펴보기: 빅데이터

—

인간에 대한 문제는 실험 데이터를 모으기 거의 불가능합니다. 관측 데이터를 이용하여 인과관계를 알아내는 수밖에는 없습니다. 일반적으로 관측 데이터로는 인과관계를 완벽하게 알 수 없지만, 빅데이터를 통해 어느 정도 살펴볼 수 있습니다.

예를 들어, 영재 고등학교가 일반 고등학교에 비해 교육의 질이 높은지 알고 싶습니다. 이 문제가 어려운 점은 학교별로 학생의 수준이 많이 다르기 때문입니다. 우리나라의 영재고는 중학교에서 수학・과학에 뛰어난 재능을 보여준 학생을 선발합니다. 영재고 학생이 수능에서 좋은 성적을 얻고 좋은 대학에 진학하는 것은 자명해 보입니다. 그러나 이러한 데이터는 영재고가 효과적인지 알려주지 않습니다. 영재고를 진학한 학생의 능력이 출중해서 좋은 대학으로 진학하는지, 아니면 영재고 교육의 질이 좋아서 높은 대학진학률을 보이는지 분간할 수 없습니다. 비슷한 수준의 학생을 대상으로 실험을 해야 교육의 질에 대한 평가가 가능합니다.

하지만 빅데이터를 이용하면 인과관계를 어느 정도 엿볼 수 있습니다. 영재고를 진학한 학생 중 입학 커트라인을 간신히 넘긴 학생과 영재고에 지원했다가 아슬아슬하게 진학에 실패해서 일반고로 진학한 학생의 고등학교 졸업 시 학업성취도를 비교하면 됩니다. 모든 데이터를 사용하는 것이 아니라 비슷한 그룹 학생의 데이터만을 사용하는 것입니다. 2014년에 이러한 분석 방법을 적용하여 미국에서 특수목적고의 효과가 전혀

없었다는 결론을 내린 논문이 경제학의 저명한 저널에 게재됩니다. 제목이 〈엘리트 환상〉이었습니다. 단, 이러한 데이터 분석은 빅데이터가 있어야 가능합니다. 커트라인과 매우 근접한 학생의 자료를 충분히 모으려면 많은 데이터가 필요하기 때문입니다.[10]

폭력적인 영화를 많이 보면 폭력 사건이 늘어날까요? 매우 논의가 많은 주장입니다. 논쟁은 무성한데 증거는 없습니다. 일반 사람을 대상으로 실험하는 것은 불가능합니다. 관측자료로 간접적으로 살펴볼 수 있을 뿐입니다. 범죄 정보와 영화 상영 정보를 결합해서 폭력 영화 관람객 수와 폭력 범죄 발생 횟수와의 관계를 보면 될 것 같습니다. 2011년 2명의 경제학자가 FBI 범죄 정보와 영화 흥행 순위, 그리고 영화의 폭력성에 대한 정보를 모아서 분석합니다. 결과는 놀랍게도 폭력 영화와 폭력 범죄는 반비례관계였습니다. 폭력 영화를 많이 보면 폭력 범죄가 줄어드는 것이었습니다. 직관에 반하는 결과입니다. 이후 좀 더 심층적인 분석을 통해서 폭력 영화의 관람이 인간의 폭력성을 해소하는 데 도움이 된다는 점이 밝혀졌습니다. 더군다나 폭력 영화를 관람할 때는 친구와 이야기하지도 않고 술도 마시지 않습니다. 폭력 영화 관람보다는 폭력적인 친구와 모여서 술을 마실 때 폭력성이 훨씬 쉽게 그리고 더 많이 발현됩니다.[11]

현재 우리나라에서 이슈가 되고 있는 모바일게임 중독에 대해서도 과학적으로 살펴볼 필요가 있습니다. 모바일게임에 중독된 사람이 모바일게임을 하지 않았다면 정상적으로 살아갔는지를 알아봐야 합니다. 원래 심리적으로 문제가 있는 사람이 모바일게임에 쉽게 중독될 수 있기 때문입니다. 그런데 사람이 인생을 2번 살 수는 없습니다. 특별한 데이터과학이 필요할 것 같습니다. 게다가 경마나 경륜 등의 도박성 스포츠는 허용하면서 모바일게임을 규제하는 것은 뭔가 이율배반적인 것 같습니다. 모바일게임 산업의 발전도 같이 고려해야 합니다. 감과 고정관념에 의한 규제가 아니라 데이터에 기반을 둔 합리적인 판단이 필요한 이유입니다.

상당한 인과관계

인과관계를 밝히는 작업은 어렵습니다. 특히 인간과 관련된 주장의 인과관계를 밝히는 것은 임의화 실험이 불가능하여 매우 어렵습니다. 하지만 인과관계 여부는 우리의 일상생활과 직접적인 연관이 있습니다. 경찰공무원의 자살에 대한 순직 여부 결정, 제조업 노동자의 백혈병에 대한 산재 인정 여부 결정, 자살한 군인의 보훈 대상 지정 여부 등, 인과관계 인정 여부가 첨예하게 대립하는 사건은 뉴스에서 쉽게 접합니다. 경찰 업무에 의한 스트레스가 자살의 원인인지, 열악한 환경이 백혈병의 원인인지, 군 생활이 자살의 원인인지를 알아야 정확한 판단이 가능합니다. 그러나 이러한 인과관계를 데이터를 통해서 직접적으로 아는 것은 불가능하고, 여러 관측 데이터를 살펴보고 정황을 조사한 후 합리적으로 판사의 양심에 따라서 판단할 수밖에 없습니다.

이 같은 첨예한 이슈에 대해서 법원은 '상당한 인과관계'라는 단어를 사용합니다. 당직실에서 숨진 전공의는 산재 인정을 받은 반면 세월호 잠수사의 골 괴사 사건은 인과관계 인정을 받지 못했습니다. 이러한 법원의 판단을 보면서 데이터과학의 한계를 생각해봅니다. 데이터를 바탕으로 모두가 동의하는 인과관계를 발견한다는 것은 매우 어렵습니다. 우리가 기술에 대한 지식뿐만 아니라 데이터과학의 한계도 잘 이해해야 하는 이유입니다.