세상과 데이터를 잇다

KBS 데이터저널리즘팀 정한진

어떤 기사

"커피전문점의 주요 고객은 여성이라는 통념을 뒤집고
남성이 여성보다 커피를 더 많이 소비하는
것으로 나타났다."

**카드가 카드 결제 정보인 빅데이터를 활용해
소비 트렌드의 변화를 분석한 결과다.

어떤 기사

**카드 측은

“여성이 커피전문점을 자주 이용하지만,

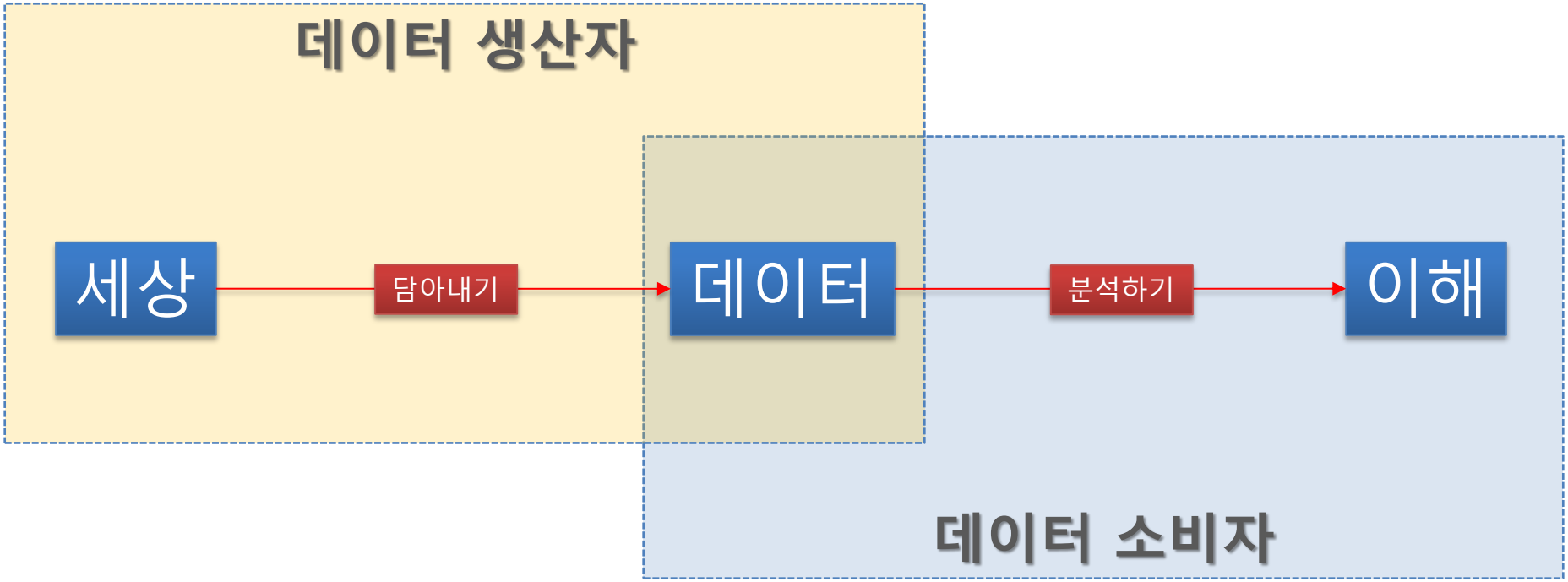
단체 자리 또는 데이트에선 남성이 커피 값을

내는 경우가 많아 남성의 커피전문점 이용 실적이

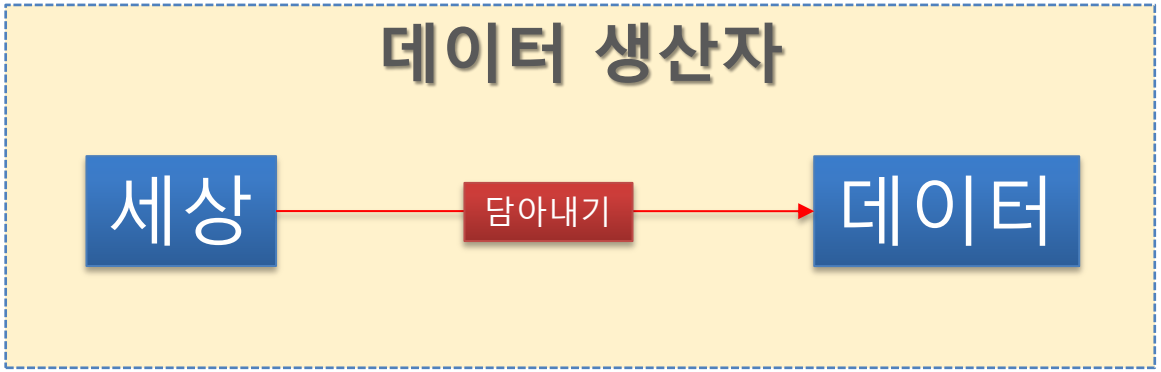
더 높은 것으로 보인다.”

고 분석했다.

세상과 데이터를 잇다



데이터로 담아내기



- ✓ 어떤 특수한 목적에 따라 데이터를 저장
  - 사건 전체가 아니라 일부만 기록
  - 결재 금액, 결재자 정보

- ✓ 사회현상 전체를 담아내지 못한다
  - 비용 문제
  - 현금 결재자
  - 원두를 직접 사서 내리는 소비자

데이터는 데이터다

데이터를 수집하는 단계에서 일어날 수 있는

여러 문제를 고려한 뒤

‘어디에 사용할 것인가’를 생각

모집단과 표본

✓ 통계에 있어서 가장 중요한 것은, 표본을 근거로 어떤 결론을 내릴 때 그 표본이 모집단 전체를 대표하는 것이라야 한다

✓ 표본은 모집단으로부터 순전히 우연에 의해 추출되어야 한다. 즉 모집단 안에 있는 개체들이 표본에 선택될 기회가 동일해야 한다.

여론조사

- ✓ 여론조사란 표본 추출에 있어 불공평한 왜곡이 형성되는 원인과의 끊임없는 싸움

- ✓ 이 전쟁에서 절대로 이길 수 없다
  - 거리에서 조사 → 집에 남아 있는 사람 제외
  - 집집마다 가정 방문 → 직장인들 제외
  - 저녁에 면접 → 영화구경을 가거나 나이트클럽에 간 친구들 제외

통계 오류 – 선택의 오류

미국-스페인 전쟁(1898년) 동안 미 해군의 전사자는 천 명당 9명이었다.

그런데 같은 기간 뉴욕시의 사망자는 천 명당 16명이었다.

- ✓ 뉴욕시민의 연간 사망률보다 해군의 연간 사망률이 낮다
- ✓ 두 집단은 비교가 불가능한 집단이다
  - ▪ 해군은 대부분 육체적으로 건강한 청년들로 구성
  - ▪ 뉴욕시민 중에는 갓난아이도 있을 것이고 노인이나 환자들도 포함되어 있어서 사망률이 높다
- ✓ 쓰레기를 넣으면 쓰레기가 나온다

적은 표본으로 실험을 하는 이유?

"OO치약으로 충치 27% 감소"

- ✓ 표본의 수를 늘릴수록 표본의 평균값은 모집단의 평균값에 가까워진다.

  (대수의 법칙)

- ✓ 실험 표본이 대규모이면 우연에 의해 나타나는 차이가 아무래도 미미해진다
  - 오차의 범위가 줄어든다
  - OO치약으로 충치 3% 감소?
- ✓ 불충분한 표본을 채택하면 순전히 우연에 의해 실험집단에서 큰 차이가 나는 결과를 얻을 수 있다
  - 제품 광고에 적극 활용

5년마다 생기는 숫자들

'소득격차 3대 지표' 모두 호전…정부 “복지확대 정책 성과”

이들 숫자가 나타내는 내용들은 어떤 주기성이

있는 것이 아니라,

매 5년마다 선거가 실시되기 때문이다

5년마다 생기는 숫자들

- ✓ 사용 전, 사용 후 사진법
  - before-after photograph
  - 잡지나 광고에서 자주 사용되는 속임수
  - 샴푸의 덕택이 아니고, 사진사의 기술 덕택

- ✓ 사전, 사후 눈속임
  - before-after trick
  - 사전, 사후를 단순히 비교하는 형식에 통계를 사용

어떤 조사 방법의 결과인가?

"나이가 많은 부인일수록 팔자걸음을 하게 된다."

- ✓ 여성의 연령과 신체적 특징 사이의 관계를 조사
- ✓ 걸음걸이에 있어 양쪽 발자국 사이의 각도를 측정

어떤 조사 방법의 결과인가?

- ✓ 한 여성을 일정기간에 걸쳐 조사해야만 한다
  - 일정 시점에서 표본 조사
  - 현재 나이든 여성들이 젊었을 때에는 팔자걸음으로 걸으라고 교육
  - 현재 젊은 여성들은 팔자걸음을 걸으면 안 된다고 배워왔다
  - 횡단조사 → 종단조사로 변경 필요

- ✓ 횡단조사
  - 일정 시점에서 광범위한 표본을 대상으로 한 번 조사
- ✓ 종단조사
  - 시간의 흐름에 따라 조사대상이나 상황의 변화를 측정
  - 일정한 시간 간격을 두고 동일한 내용을 반복적으로 측정하여 자료를 수집하거나 조사

기록의 일관성

"유행성 감기와 폐렴에 관한 최근의 통계를 보면
거의 80%가 남부의 세 주에서 발생한다"

✓ 다른 주에서는 이미 중단된 지 오래된
이 병의 발생 보고를 이 세 주에서만은
아직도 의사들이 의무적으로 시행

기록의 일관성

"1940년 이전 미국 남부에는 연간 수십만 명의
말라리아 환자가 발생했는데
오늘날에는 손으로 꼽을 정도밖에는 발생하지 않는다"

✓ 오늘날에는 말라리아라고 명확히 판명된 경우에 한해서만 보고 되는데 반해 옛날 미국 남부 대부분의 주에서는 말라리아라는 단어가 감기나 몸살을 나타내는 일상용어로서 사용되었다

통계 조작의 함정

특정 학군에 속한 고등학생들의 시험 성적과 졸업률

- ✓ 교육 담당 공무원 평가 및 보상 기준
- ✓ 좋은 평가와 보상을 받는 비도덕적 방법
  - 성적이 가장 나쁜 학생들이 시험을 치르지 못하게
  - 졸업 전에 학교를 떠나는 학생들을 전학으로 처리(중퇴→전학)
- ✓ <u>주어진 목표와 동떨어진 방법을 통해 통계적으로 더 나아 보이게 만드는 경우</u>

통계 조작의 함정

심장동맥 확장 수술을 받은 환자들의 사망률을 측정

- ✓ 심장병 전문의들에 대한 평가
- ✓ 일반 소비자는 이러한 정보에 접근할 수 없다
  - ▪ 정부에서 이러한 데이터를 수집하여 알리는 것
- ✓ 의사가 환자의 사망률을 낮추는 가장 쉬운 방법
  - ▪ 심각한 병세를 보이는 환자들의 수술을 거부하는 것
  - ▪ 의료 행위에 대한 의사들 개개인의 판단에 영향을 미친다

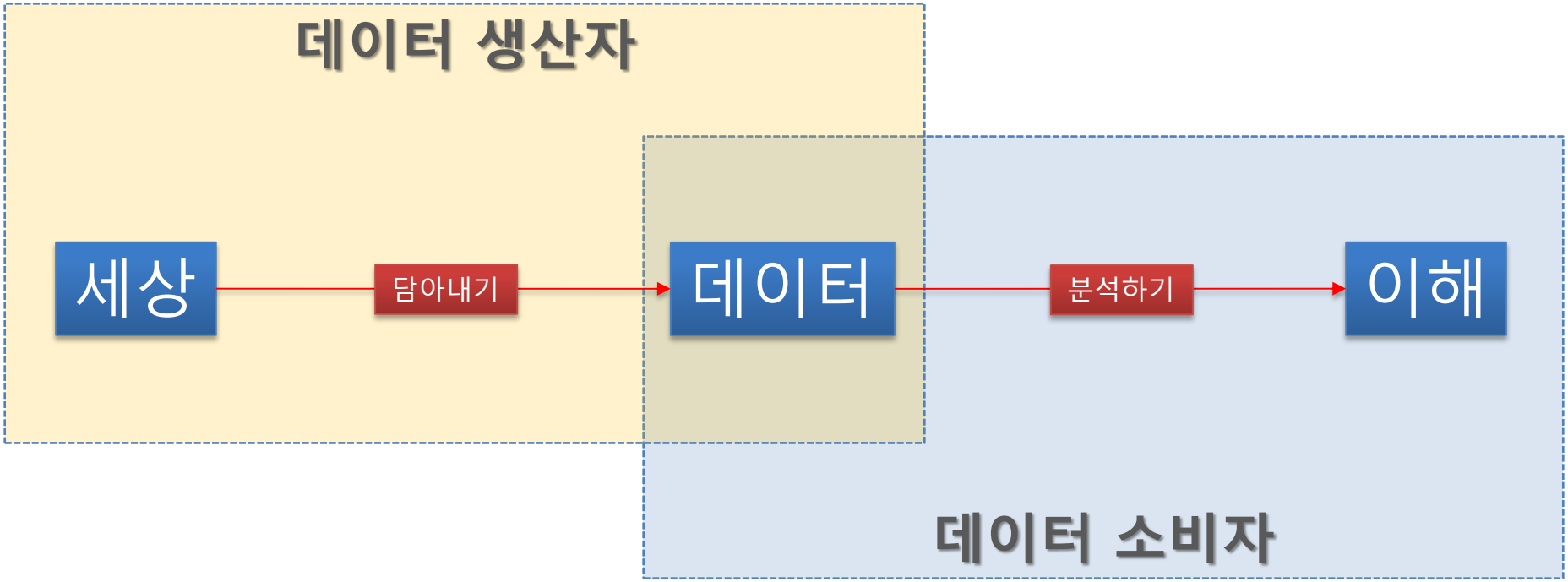
빅데이터 여론 조작

“복잡한 현실보다 데이터 자체를 조작하기가 더 쉽다”

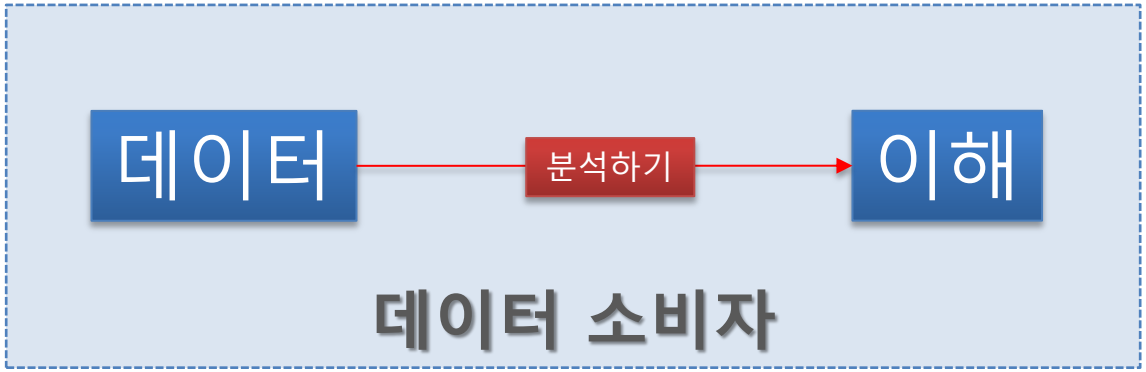
✓ 현실을 바꾸기 보다는 현실을 대리하는 데이터를 조작

- 쉽게 말해 장난치기가 쉽다
- 2018년 네이버 댓글 조작
- 2012년 대통령 선거 트위터 여론 조작

세상과 데이터를 잇다



데이터 분석하기



✓ 19세기 영국 수상 벤자민 디즈레일리

"세상에는 3가지 거짓말이 있다.
그럴 듯한 거짓말, 새빨간 거짓말, 그리고 통계다.
그 중 통계는 오용의 위험성이 있다."

속지 않으려면 통계학을 배워라

데이터를 사용해 예상이나 분석을 하는데

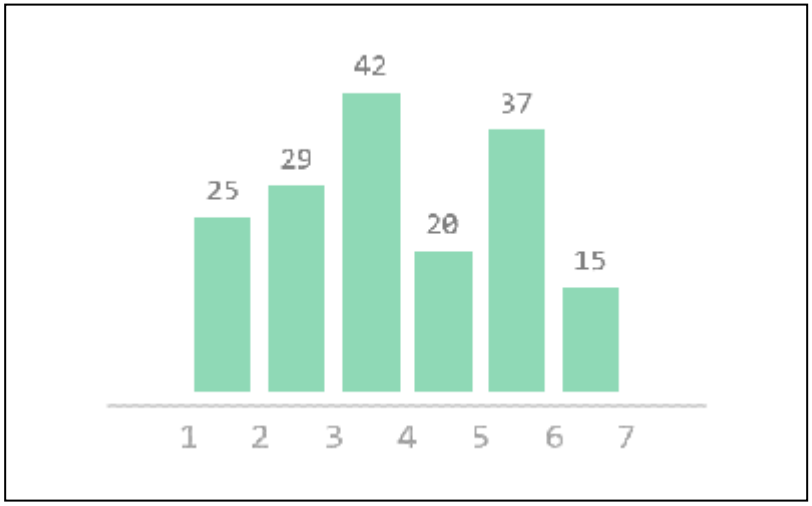
이때 쓰이는 각 수치가 어떻게 만들어지는지

또 그 수치에는

문제점은 없는지 알아야 한다

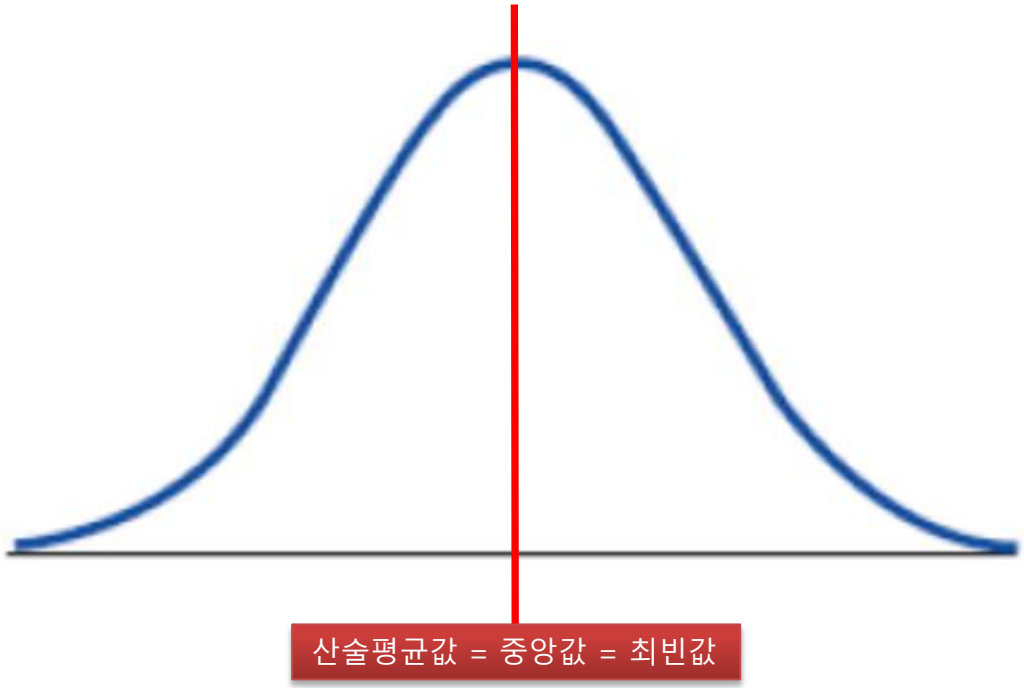
대푯값에 의한 정보축약

- ✓ 대푯값 : 산술평균값, 중앙값, 최빈값
- ✓ 각 값들의 정의를 알고 그 값들 간의 차이도 알아야 한다
- ✓ 분석과정에 '도수분포표'를 꼭 그려보자



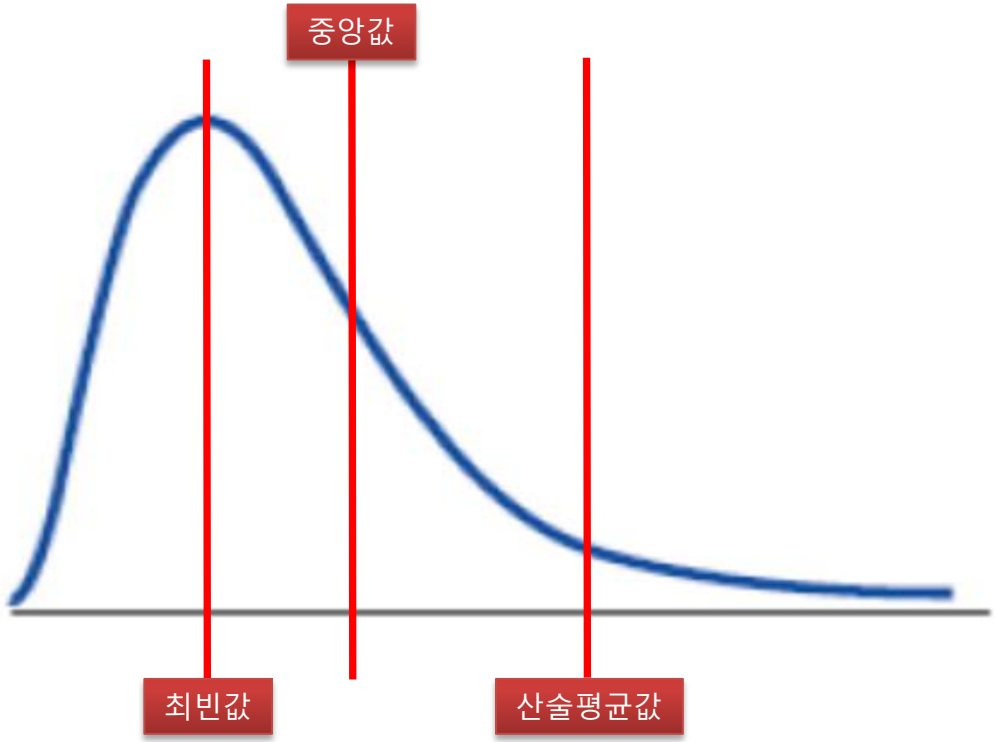
대푯값에 의한 정보축약

- ✓ 신체의 체위(키, 몸무게 등)는 정규분포를 따른다
- ✓ 산술평균값 = 중앙값 = 최빈값



대푯값에 의한 정보축약

- ✓ 소득의 분포는 왼쪽으로 치우친 그래프
- ✓ 최빈값 < 중앙값 < 산술평균값
- ✓ 대푯값이 여러 개로 나타날 수 있다



소득을 이야기 할 때

- ✓ 소득의 분포
  - 최빈값 < 중앙값 < 산술평균값
- ✓ 산술평균값은 소득 하위계층과 소득 상위계층 그 어느 쪽도 해당하지 않는 터무니없이 황당한 수치
- ✓ 한 가정의 소득이 가족 수에 비례한다는 가정
  - 즉 4인 가족의 소득이 2인 가족의 소득의 2배가 되는 것은 쉽지 않다
- ✓ 어느 해의 '평균값'(산술평균값)과 다른 해의 '평균값'(이번에는 중앙값)을 비교하는 것은 아무 의미가 없다

산술평균값과 중앙값의 함정

- ✓ 만약 산술평균값과 중앙값이 차이가 나지만 이중 하나만을 나타낸다면
  - 간결함을 위해서거나
  - 누군가가 통계를 이용해 '설득'을 하고자 하기 때문이다

산술평균값과 중앙값의 함정

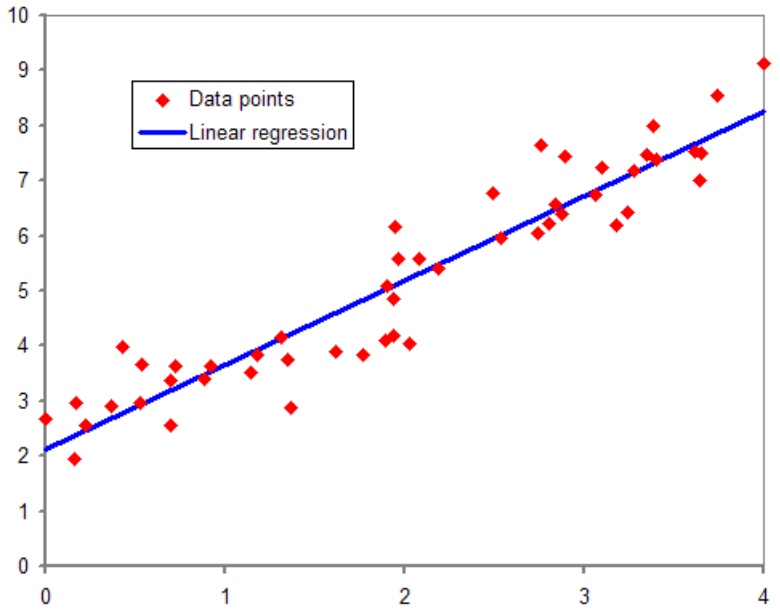
- ✓ 치명적인 병에 걸렸다
- ✓ 효과적일 수 있는 신약이 개발되었다
  - 같은 병에 걸린 환자의 기대 수명의 중앙값이 2주 늘어난다
  - 많은 환자들이 효과를 보진 못했지만

    적지 않은 수의 환자(10 ~ 15%)가 완치되었다
- ✓ 이 경우 대푯값으로부터 이탈한 값이 당신의 결정에

  큰 영향을 줄 수 있다

회귀분석

- ✓ 데이터와 데이터 사이의 관계를 발견
- ✓ 관계를 정량화하고 그 영향과 효과를 양적으로 알 수 있다
- ✓ 분석과정에 '산포도'를 꼭 그려보자

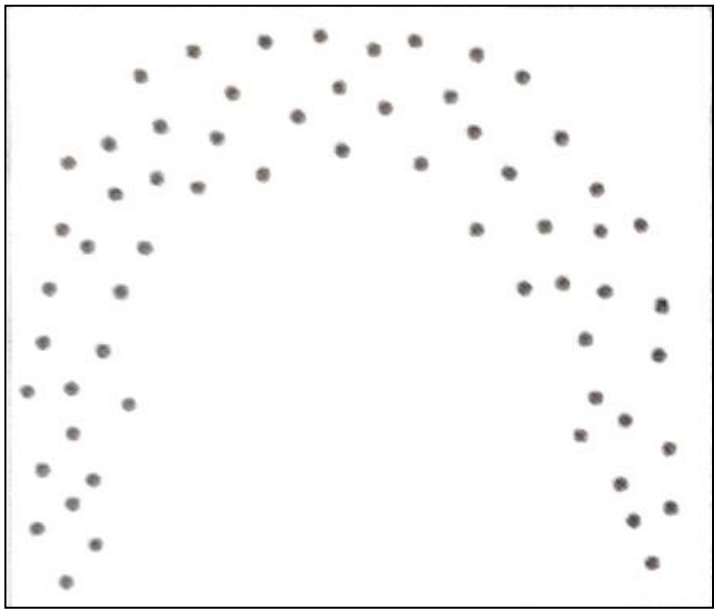


회귀분석

- ✓ 회귀식과 표준편차는 변수 간의 관계를 축약
  - 원래 데이터의 성질 일부를 생략하는 것에 지나지 않는다
  - 산포도를 회귀식에 더해 표기하자
- ✓ 회귀식은 인과관계를 발굴하는 방법
  - 학술연구에서는 '인과관계'의 발견이 최종목표
- ✓ 회귀분석이 나타내는 것은 '상관관계'이지 '인과관계'가 아니다
  - 상관관계 : 'a'와 'b'가 동시에 관찰된다
  - 인과관계 : 'a'에 의해 'b'가 일어난다

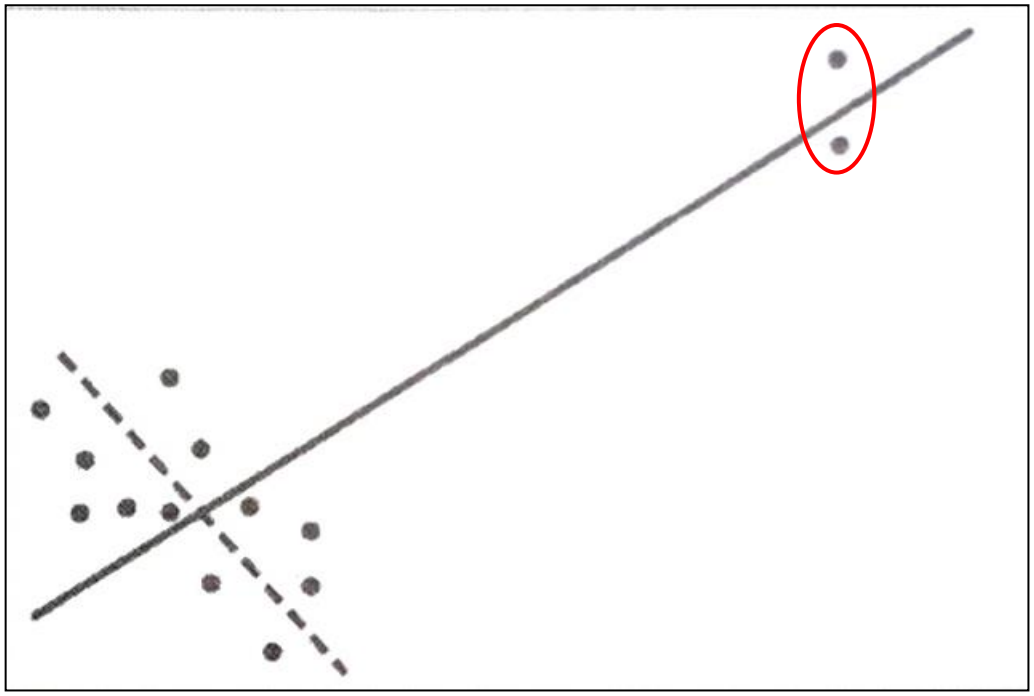
산포도를 그리지 않으면 함정에 빠질 수 있다

✓ 직선적인 관계는 없지만 관계가 없다고 이야기할 수는 없다



산포도를 그리지 않으면 함정에 빠질 수 있다

- ✓ 불과 2개의 벗어난 값이 말도 안 되는 방향으로 바꾼다
- ✓ 데이터 개수가 충분할 때 종속변수의 상위 1%와 하위 1%를 버린다

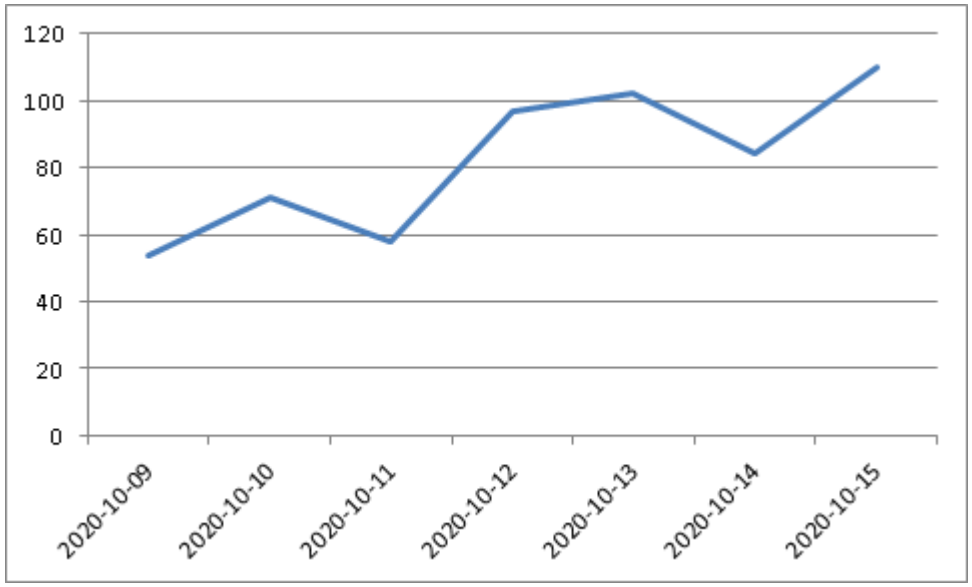


시계열 분석

- ✓ 회귀모델의 목적은 인과관계의 탐구
  - 다양한 행동의 결과를 예측 : 키와 몸무게
  - 회귀분석은 실행하기 전에 무엇을 독립변수로 선택할 것인지가 큰 문제(주관적 선택)

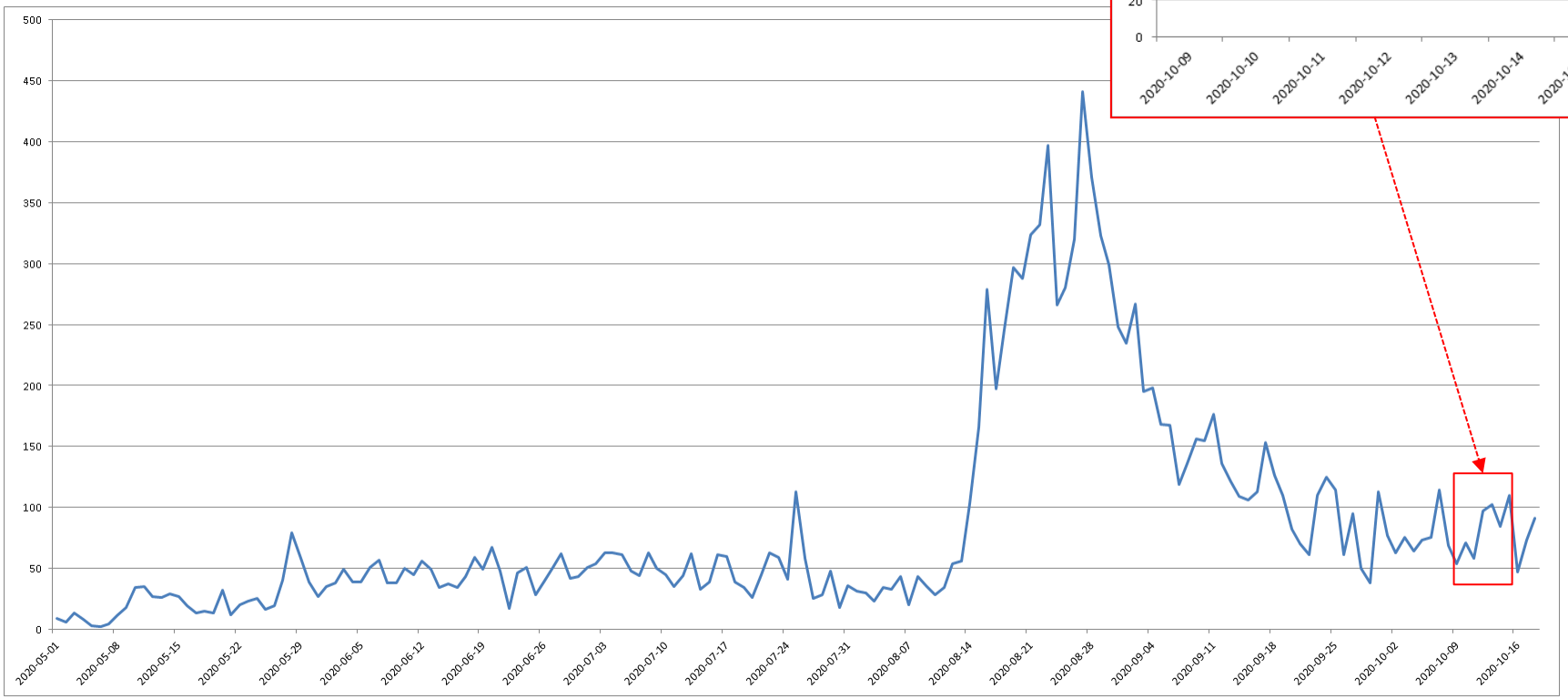
- ✓ 시계열 분석의 목적은 최소의 노력으로 최적의 예상을 하는 것
  - 다양한 변화를 예상 : 내일의 주가는?
  - 어떤 데이터의 과거 상태에서 미래를 예상한다
  - 특히 단기적인 예상을 할 때 큰 힘을 발휘
  - 오늘의 결과값에는 무수한 요인에 관한 정보가 축약되어 있다
  - 내일의 결과값에 대한 독립변수로서 '오늘의 결과값'을 이용해 하나의 독립변수를 가진 회귀를 실행(유용성이 높은 단순화)

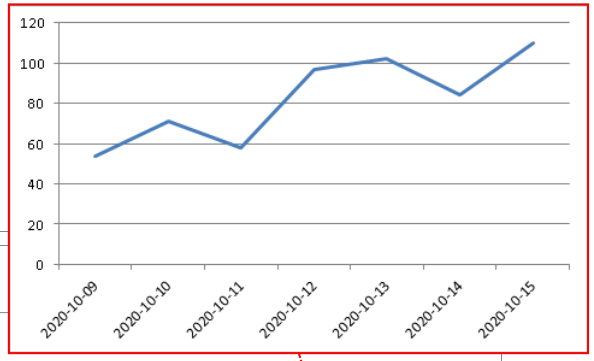
어떤 그래프


120
100
80
60
40
20
0
2020-10-09
2020-10-10
2020-10-11
2020-10-12
2020-10-13
2020-10-14
2020-10-15

통계 오류 – 심슨의 역설

✓ 부분에서 수집된 자료와 전체의 자료가 거의 반대의 결과를 낳는 오류





인과관계와 상관관계의 구분이 왜 중요?

- ✓ 상관관계 : ‘a’와 ‘b’가 동시에 관찰된다
- ✓ 인과관계 : ‘a’에 의해 ‘b’가 일어난다
- ✓ 인과관계에 근거해서 돈과 시간을 원인에 해당하는 정확히 투자하면 좋은 결과를 얻을 확률이 높다

인과관계 or 상관관계 ?

美 **의학전문지**

"초콜릿 많이 먹는 나라가 노벨상 수상자도 많아"

- ✓ **초콜릿 소비량이 많아서 노벨상 수상자가 많다(인과관계)**
- ✓ **노벨상 수상자가 많은 나라들은 초콜릿 소비가 많다(상관관계)**

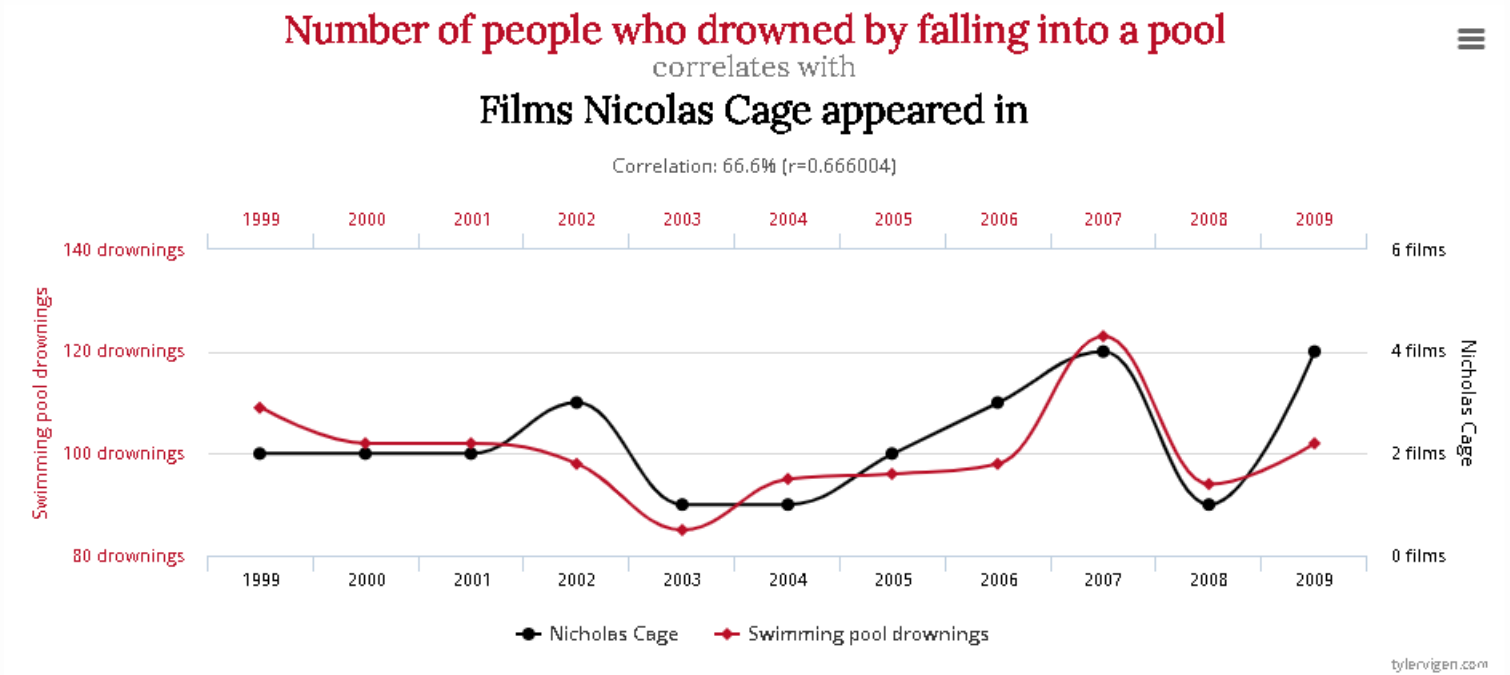
인과관계 or 상관관계 ?

두 변수의 관계가 인과관계인지 아니면 상관관계인지를 확인하기 위해서는 다음 사항들을 의심해봐야 한다.

- ✓ '우연의 일치'는 아닌가?
- ✓ '제3의 변수'는 없는가?
- ✓ '역의 인과관계'는 존재하지 않는가?

우연의 일치 = 거짓 상관

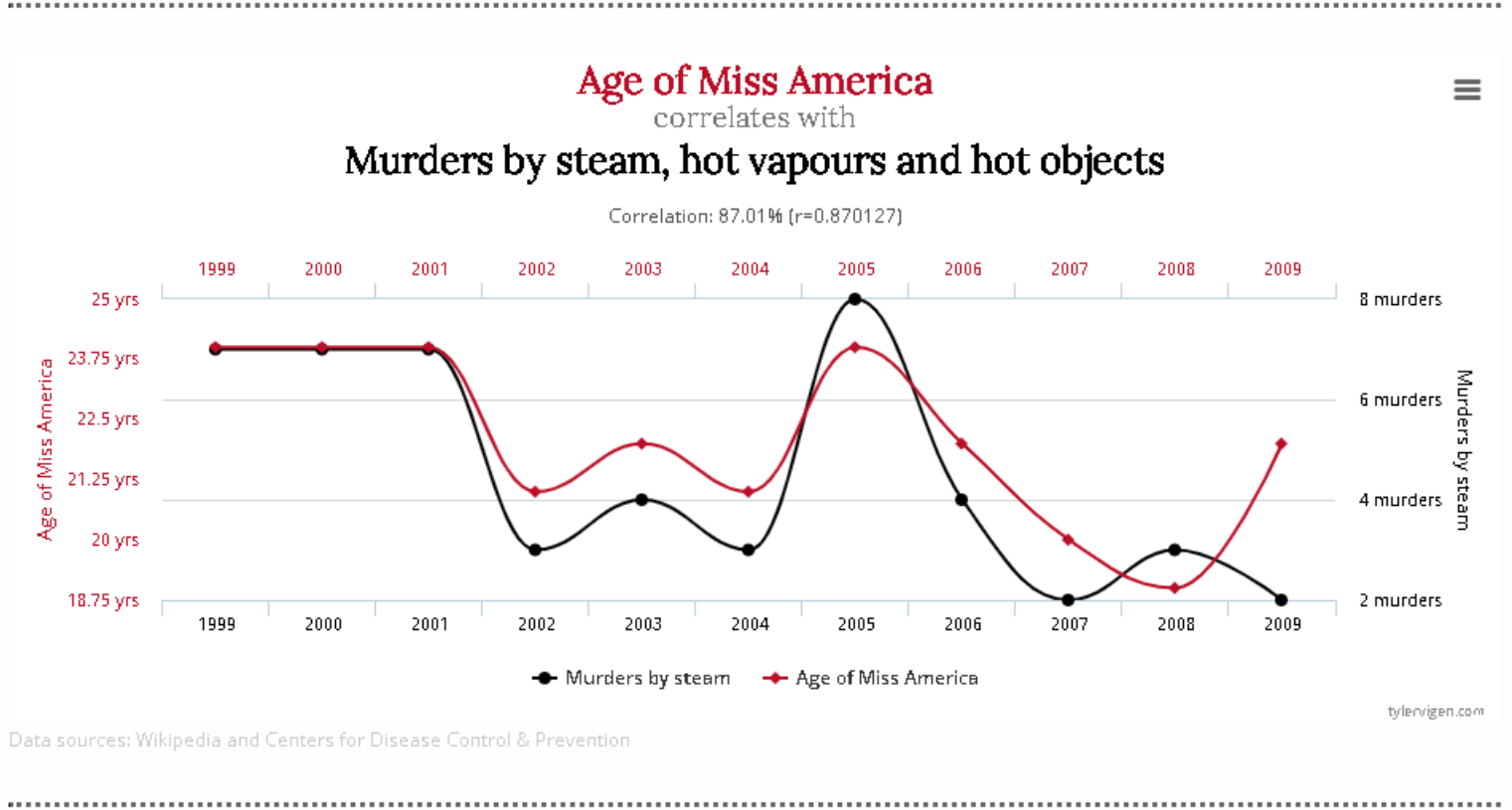
✓ 수영장 익사자 수와 니콜라스 케이지의 연간 영화 출연 편수



http://tylervigen.com/spurious-correlations

우연의 일치 = 거짓 상관

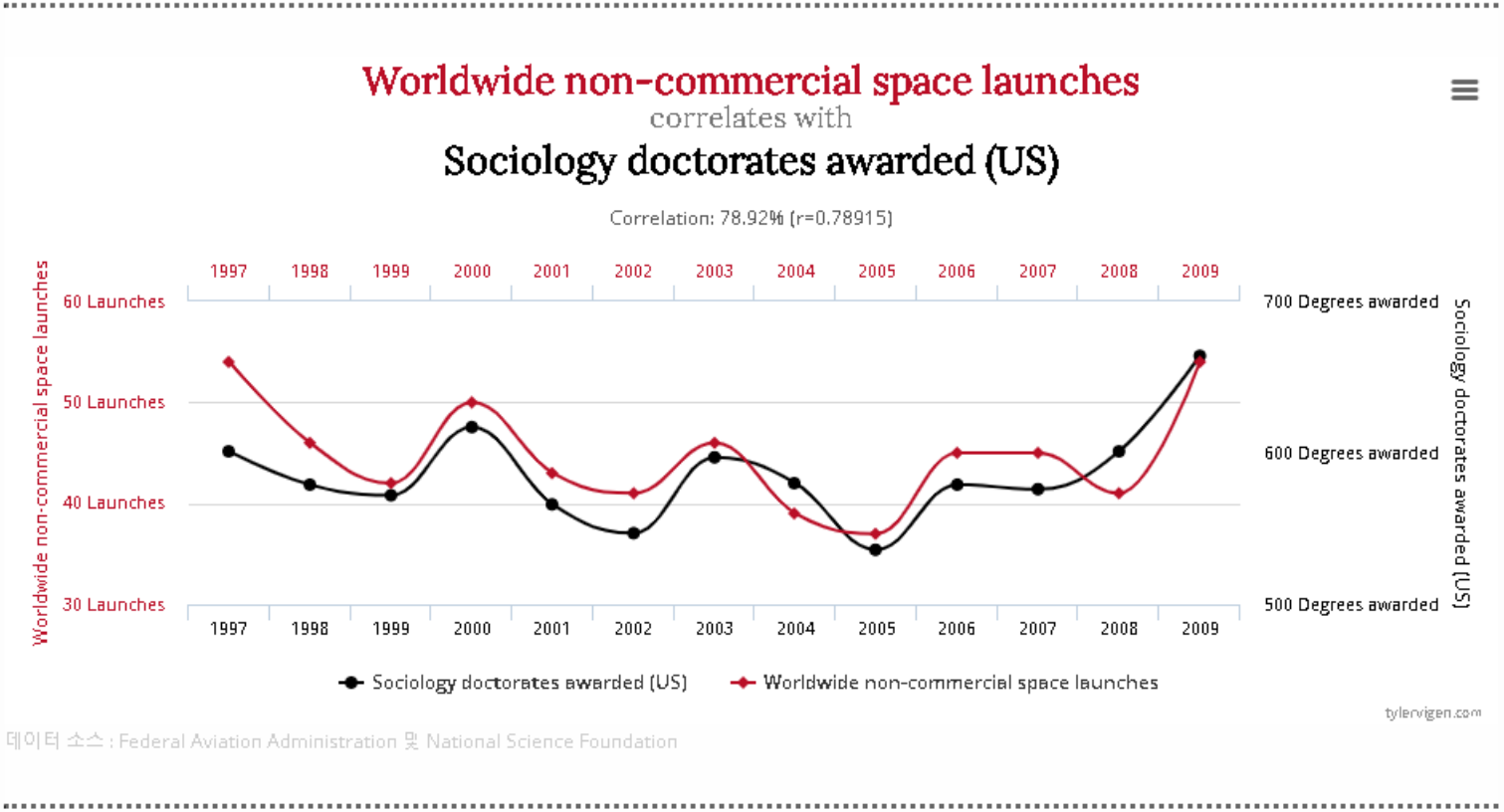
✓ 미스 아메리카의 나이와 난방 기구로 인한 사망자 수



http://tylervigen.com/spurious-correlations

우연의 일치 = 거짓 상관

✓ 전 세계 비상업적 우주 발사 수와 사회학 박사 학위자 수

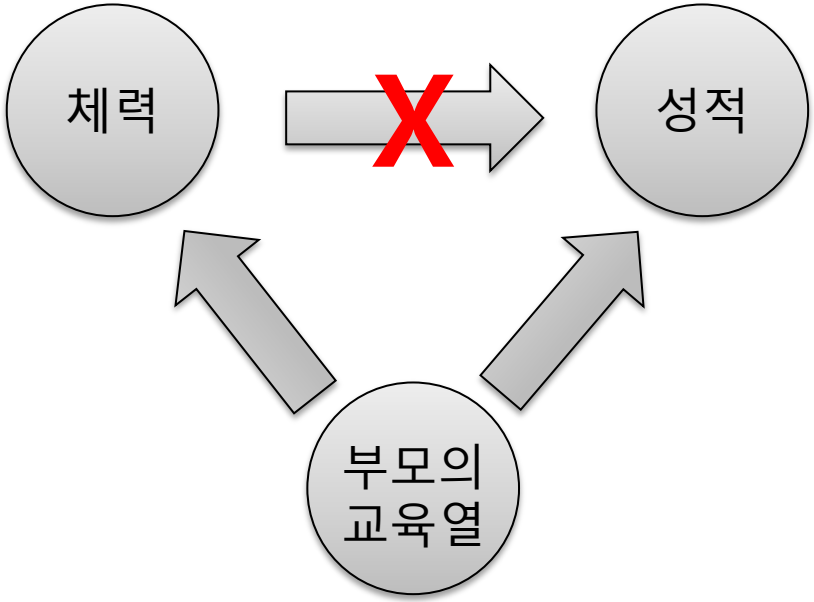


http://tylervigen.com/spurious-correlations

제3의 변수 = 교란 요인

- ✓ 원인과 결과에 모두 영향을 주는 '제3의 변수'의 존재(교란 요인)
- ✓ 상관관계에 지나지 않는 것을 마치 인과관계가 있는 것처럼 보이게 만드는 성가신 존재

체력이 좋은 아이들이 성적도 높다



제3의 변수 = 교란 요인

"메사추세츠주의 어느 장로교 목사의 수입과
하바나의 럼주의 가격 사이에는
높은 상관관계가 성립한다."

✓ 제3의 요인

- 모든 물가나 가격수준은 시간이 지나면서 전 세계적으로 상승한다

역의 인과관계

"섬 주민들은 자신들의 몸에 기생하는 이가 건강을 가져다 준다고 믿는다.
지난 수백 년 동안 관찰한 결과, 건강한 사람에게는 이가 많지만
아픈 사람에게는 종종 나타나지 않는다."

- ✓ 남태평양 뉴 헤브리디즈 섬
  - 섬 주민들 대부분에게는 항상 이가 있었다
  - 그러데 이 이로 인해 열병에 걸려 체온이 올라가면 이는 환자의 몸을 떠나게 된다
- ✓ 원인과 결과가 뒤죽박죽 뒤엉켜

역의 인과관계

- ✓ 경찰관 수와 범죄의 관계
  - 경찰이 많아서 범죄의 발생 건수가 많다?
  - 범죄가 많은 우범 지역이어서 경찰을 많이 배치했다

제3의 요인 or 역의 인과관계

- ✓ 담배를 피우면 공부를 못한다?
  - 조사 결과는 그렇다는 것으로 판명되었다
  - 흡연과 성적불량이 동시에 발견
  - 흡연이 성적불량의 원인이라는 부당한 엉터리 가정을 하였던 것
  - 전후 관계와 인과관계를 혼동하는 오류
- ✓ 어쩌면 성적이 불량하기 때문에 흡연을 하게 된 것인지도 모른다(역의 상관관계)
- ✓ 아님 양쪽 모두에 영향을 주는 제3의 요인(교란 요인)의 결과라고 결론을 내릴 수도 있다

인과관계 증명 : 실험군과 대조군

원인이 발생한 경우(실험군)의 결과

원인이 발생하지 않은 경우(대조군)의 결과를

비교해서 인과관계가 존재함을 증명

인과관계 사례

의료비 본인 부담률과 사람들의 건강 상태 사이에 인과관계가 있을까?

- ✓ 랜드 의료보험 실험
  - 현재 가치로 무려 3억 달러의 연구비 쏟아부음
  - 지출을 줄이기 위해 병원에 내원하는 걸 꺼리게 되고 결국 질병의 조기 발견 및 치료 시기를 놓쳐 건강상태가 악화될 우려

- ✓ 의료비 본인 부담률과 사람들의 건강 상태 사이에 인과관계가 없다
  - 단, 소득이 낮고 건강 상태가 나쁜 사람들의 경우 본인 부담률 인상이 건강 상태를 악화시키는 결과

인과관계 사례

어린이집을 늘려도 어머니의 취업률은 상승하지 않는다

- ✓ 사적 보육 서비스 → 공적 교육 서비스로 갈아타는데 일조
- ✓ 오히려 어린이의 발달과 건강에 긍정적인 영향
  - ▪ 어린이집의 전문적인 지식과 기능을 갖춘 보육 교사

인과관계 사례

'최저임금'과 '고용' 사이에 인과관계는 있을까?

- ✓ 최저임금 상승은 고용을 감소시키지 않는다
- ✓ '경기 악화'는 최저임금과 고용 모두에 영향을 미치는 '교란 요인'
  - 기업들은 최저임금의 상승에 따른 비용 부담을 구조조정이 아니라 가격에 반영해 타개하려고 함
  - 물가 상승 초래
- ✓ 최저임금을 규정하는 법안의 요점
  - 저임금 노동자들에게 최소한의 소비가 가능한 시급을 보장
  - 과거에 비해 액수만 커지고 실제로 살 수 있는 물건은 오히려 줄어드는 시급을 지급하기 위한 것이 아니다

인과관계 사례

엄마, 아빠 말 안 들으면 경찰 아저씨가 잡아간다

- ✓ 스케어드 스트레이트 교육법
  - 아이들에게 두려움을 느끼게 해 왜 올바른 행동을 해야 하는지를 가르치는 방법
  - 1970년대 미국의 한 텔레비전 프로그램에서 이 교육법을 체험한 청년 그룹이 더 이상 범죄에 연루되지 않았다고 보도
- ✓ 전후 비교설계를 안이하게 이용해 잘못된 결론에 이른 전형적인 사례
  - 이후 실험군과 대조군으로 나눠 비교
  - 오히려 그들을 범죄자로 만들 확률을 높이고 있었다는 결과

인과관계 사례

공부 잘하는 친구와 사귀면 성적이 오를까?

- ✓ 또래집단 효과
  - 공부 잘하는 친구와 사귀기 때문에 아이의 성적이 높아지는 것일까? (인과관계)
  - 성적이 높은 아이일수록 공부 잘하는 친구와 사귀는 것일까? (상관관계)
- ✓ 공부 잘하는 친구들에 둘러싸여도 성적은 오르지 않는다

인과관계 사례

명문대를 졸업하면 연봉이 높을까?

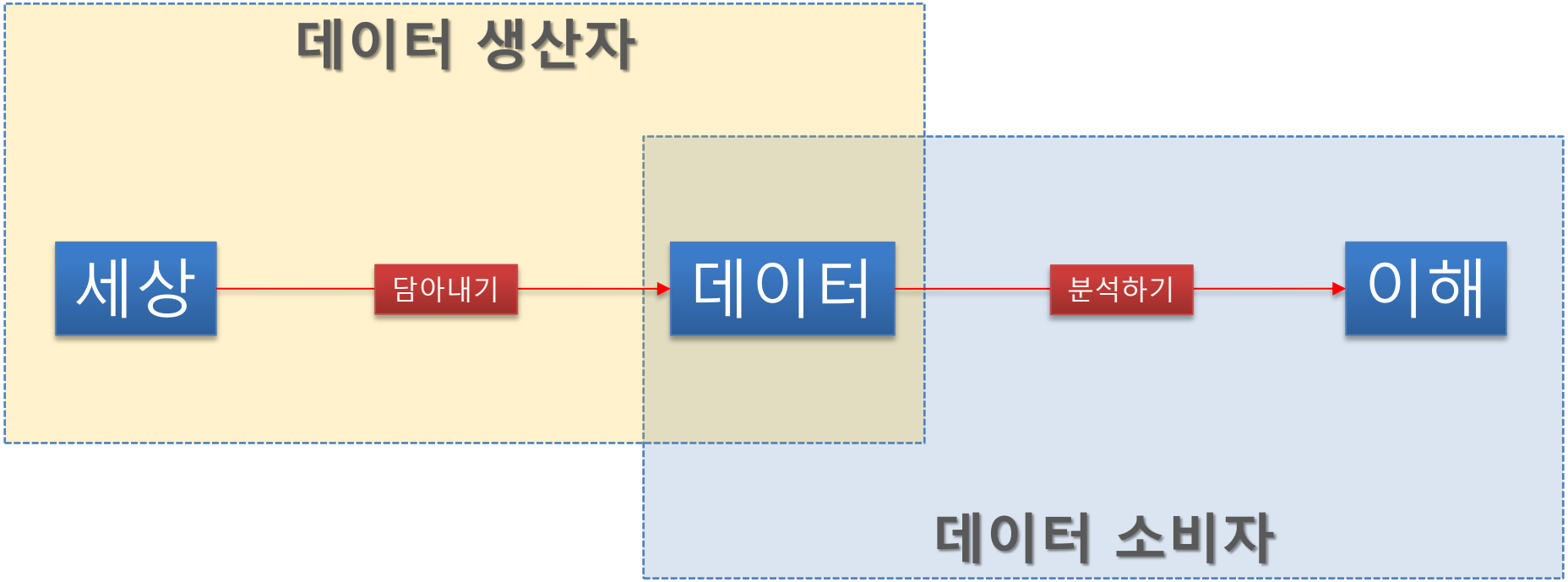
- ✓ 출신 대학이 미래 수입에 미치는 효과
  - 입학 점수가 높은 대학에 갔기 때문에 수입이 높은 것일까? (인과관계)
  - 잠재 능력이 높아 수입이 높은 업종에 취직하는 사람이 입학 점수가 높은 대학을 선택하는 것일까? (상관관계)
- ✓ 유의미한 차이는 없었다
  - 미국의 경우 표준점수가 높은 대학을 통해 구축되는 인적 네트워크가 인종적 소수자나 빈곤층에 유리하게 작용

연구결과를 기사로 작성할 때

'게젤의 준거'에 관한 문제

- ✓ 아놀드 게젤(Arnold Gesell, 1880 ~ 1961)
  - 아동 발달 분야 발전에 기여한 심리학자이자 소아과 의사
- ✓ 표준값와 자기 아이와의 근소한 수치 차이가 부모들의 고통을 유발하는 현상
- ✓ 연구자가 연구결과를 발표했을 때
  - 기자가 기사를 작성하면서 독자들이 알아야 할 정말로 중요한 숫자 몇 개를 빼놓고 전달하기 때문
  - 평균값 이외에 '정상적인 것'의 분포의 범위를 동시에 표기

세상과 데이터를 잇다



과학적 사고법

- ✓ 통계적 사고에 가까운 귀납적 추론
  - 지금까지 관찰하고 경험해온 구체적이면서도 개별적인 사례들(데이터)로부터 가설을 설정하고 실험 결과로 가설을 정당화하는 과정
- ✓ 논리의 검증에 필요한 연역적 사고
  - 귀납법에 의해 정당화된 전제를 통계적으로 타당한지 아닌지 검증하기 위해 실험을 설계하고 결과를 예측하는 과정
- ✓ 현대적인 과학적 사고법
  - 귀납법에 의해 정당화된 전제를 사용해 연력적으로 사고하는 것

감사합니다.

KBS 데이터저널리즘팀 정한진

bururu@kbs.co.kr

참고자료

1. 대럴 허프 『새빨간 거짓말, 통계』(더불어책, 2003)

2. 이다 야스유키 『통계학 리스타트』(비지니스맵, 2010)

3. 찰스 윌런 『벌거벗은 통계학』(책읽는수요일, 2013)

4. 캐시 오닐 『대량살상 수학무기』(흐름출판, 2017)

5. 나카무로 마키코 · 쓰가와 유스케 『원인과 결과의 경제학』(리더스북, 2018)

6. 한스 로슬링 『FACTFULNESS』(김영사, 2019)

7. 타일러 비겐 『Spurious Correlations』(http://tylervigen.com/spurious-correlations)