데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

재현성 위기와 연구윤리

1. 대부분의 연구결과는 거짓이다?

유리겔라를 아시나요?

- 본인이 초능력자라고 주장하는 유리겔라는 1984년 방한 당시 TV 프로그램에서 손을 대지 않고 숟가락을 구부리는 마술(이라고 쓰고 속임수라고 읽는다)을 선보여 전국민을 충격의 도가니에 빠트렸다(관련 영상은 유튜브에서 찾아볼 수 있다).

초능력자는 존재하는가?

- 초능력자(초감각 인지능력 소유자)에 관한 이야기는 과학계에서도 심심찮게 등장하는데 2011년 미국의 저명한 사회심리학자 대릴 벰 교수는 다음과 같은 실험 결과를 통해서 초감각 인지능력이 존재한다고 주장하였다.
- 우선 100명의 학생들에게 두 개의 커튼을 보여주는 컴퓨터 스크린 앞에 앉아서 왼쪽이나 오른쪽 커튼 중 어디에 이미지가 숨어 있는지를 고르게 했다. 그런 다음 커튼을 젖혀서 정답 여부를 알려주었다. 이런 과정을 36개의 이미지에 대해서 실시하였다.

초능력자는 존재하는가?

- 피실험자들은 알지 못했지만 여기서 이미지는 피실험자의 선택 이후에 임의로 배정되었다. 따라서 피실험자가 이미지의 위치를 알아 맞춘다는 것은 **예지능력**이 있는 것으로 간주되었다.
- 예지능력이 없다면(즉 귀무가설하에서) 이 위치를 맞출 확률은 50%이다. 뱀의 논문에서 피실험자들은 에로틱한 이미지가 보일 때 53%의 적중률을 보였다($p=0.01$). 이 논문에서는 이 외에도 추가로 8개의 유사한 실험을 1,000명의 피실험자들에게 10년에 걸쳐서 실시한 후 총 9개의 연구 중 8개가 통계적으로 유의미한 예지능력을 보여준다는 결과를 제시하였다.
- 그렇다면 초능력은 정말 존재하는 것일까?

대부분의 발표된 연구결과는 거짓이다?

- 앞에서 언급한 “대부분의 발표된 연구 결과는 거짓이다”라는 스탠포드 의대 존 이오아니디스 교수의 논문을 실제로 읽어본다면 이 논문은 사실 재현성 위기에 관한 경고라는 것을 알 수 있다.
- 재현성 연구란 똑같은 연구를 다른 자료를 사용했을 때 비슷한 결과가 관측되는지 알아보는 것을 말한다.
- 기존 연구들의 재현성 여부를 확인하기 위해 일련의 과학자들이 재현성 프로젝트(Reproducibility Project)라는 이름하에 100개의 심리학 연구를 더 많은 표본을 가지고 실제 똑같은 결과가 나오는지 여부를 확인하였다.

대부분의 발표된 연구결과는 거짓이다?

- 원 실험에서는 대부분의 연구(97개)에서 통계적으로 유의미한 결과가 나왔지만 재현된 연구에서는 통계적으로 유의미한 결과는 36개가 나왔다.
- 이 결과를 가지고 재현이 제대로 되지 않은 연구가 64개라는 자극적인 보도들이 쏟아졌다.
- 하지만 통계적 유의미의 기준은 p값이 0.05이하인지 여부만 이야기하는 것이고 실제 p값이 0.049에서 0.051로 바뀐 경우라면 그 차이는 미미할 수 있다. 즉 통계적으로 유의미한 연구 결과와 통계적으로 유의미하지 않은 연구 결과의 차이가 통계적으로 유의미하지 않을 수 있다.

대부분의 발표된 연구결과는 거짓이다?

- 기존의 연구 결과와 새 연구 결과의 차이가 통계적으로 유의미한지를 확인해본 결과 23개의 연구에서 기존의 결과와 새 연구 결과가 통계적으로 유의미한 차이를 보인다는 것을 알 수 있었다.
- 통계적으로 유의미한지 여부와 더불어 효과크기(effect size)와 효과의 방향 또한 확인하는 것이 중요하다.
- 많은 경우 재현연구에서 효과의 방향은 일치했지만 효과 크기는 원래 효과 크기의 절반 정도에 그쳤다. 이 얘기가 의미하는 것은 운 좋게 큰 효과가 나온 연구가 논문으로 나오는 편향을 보여주는 것으로 이러한 현상을 **귀무가설로의 회귀**라고 한다.

어디서부터 잘못된 걸까?

- 통계분석은 PPDAC의 각 단계마다 잘못될 수 있다. 먼저 문제(Problem) 자체가 데이터를 이용해서 대답할 수 없는 경우를 생각해 볼 수 있다. 예를 들면 “지난 10년 동안 영국에서 10대의 임신율이 그토록 낮아진 이유는 무엇인가”라는 질문에 관측된 데이터로는 어떤 설명도 제공해 주지 못한다.

어디서부터 잘못된 걸까?

- **계획(Plan)단계에서 다음과 같은 사례를 고려할 수 있다.**
  - 대표성이 부족한 편리하지만 저비용의 표본을 선택한 경우: 예를 들면 자동응답 전화 여론조사를 통한 선거 예측
  - 설문조사에서 오해를 불러일으킬 수 있는 문구를 사용한 경우: “온라인 구매를 통해 얼마나 절약할 수 있는가”와 같은 질문을 통해 긍정적인 답변 유도
  - 표본 크기가 너무 작아서 검정력이 낮은 연구를 설계한 경우: 대립가설이 참이라도 귀무가설을 기각하지 못할 수 있다.

어디서부터 잘못된 걸까?

- **데이터(Data)** 수집 단계에서는 응답누락, 중도 포기한 피실험자, 실험 참가자 모집부진, 모든 것을 효율적으로 코드화 하는 문제 등이 발생할 수 있다.

어디서부터 잘못된 걸까?

- **분석(Analysis)이 잘못되기 가장 쉬운 경우는 자료입력에서 실수를 하는 경우이다.**

  ⮡ 하버드 대학의 저명한 경제학자 라인하트와 로고프 교수는 2010년 긴축재정 관련 논문을 발표하였는데 이 논문을 근거로 당시 하원을 장악한 공화당에게 대대적인 복지정책 축소를 추진하였다. 하지만 이후 다른 학교 대학원생이 이 논문의 자료를 재분석하는 과정에서 원논문에서 스프레드시트 실수로 마지막 다섯 개의 자료가 분석에 포함되지 않았음이 밝혀진다. 그뿐만 아니라 이 자료를 포함할 경우 논문의 주요결론이 바뀌는 것으로 드러난다.

어디서부터 잘못된 걸까?

- **분석(Analysis)이 잘못되기 가장 쉬운 경우는 자료입력에서 실수를 하는 경우이다.**

  ⮡ 2009년 세계적인 투자회사 악사 로젠버그 사에서 일하는 프로그래머의 통계모형에 관한 코딩 실수로 고객들에게 2억 1,700만 달러의 손실을 입히는 일이 발생한다. 이 회사는 고객들의 손실보전과 벌금 2,500만 달러를 지불하였다.

어디서부터 잘못된 걸까?

- **또 다른 분석(Analysis) 단계에서 많이 저지르는 실수는 잘못된 통계모형을 적용하는 경우로 다음과 같은 경우를 고려할 수 있다.**

  ↳ “가정의학과 환자”와 같이 그룹 전체를 하나의 표본 단위로 간주해서 임의로 뽑는 경우에서 마치 개개인 환자가 임의로 표본으로 뽑힌 것처럼 간주하는 경우

어디서부터 잘못된 걸까?

- **또 다른 분석(Analysis) 단계에서 많이 저지르는 실수는 잘못된 통계모형을 적용하는 경우로 다음과 같은 경우를 고려할 수 있다.**

  → “유의미하지 않음”을 “영향 없음”과 동일시하는 경우. 즉 귀무가설을 기각하지 못한 것을 귀무가설이 참이라고 생각하는 경우. 예를 들면 술과 알코올에 관한 연구에서 일주일에 알코올을 150~200㎖ 마시는 50~64세 남성들은 사망위험이 유의미하게 줄어든 반면 그보다 덜 마시거나 더 마시는 그룹에서는 사망위험이 유의미하게 줄어들지 않았다. 하지만 사망위험의 신뢰구간을 살펴보면 이 그룹들 간의 사망위험의 차이는 유의미하지 않았다.

대부분의 발표된 연구결과는 거짓이다?

- **결론(Conclusion) 단계에서 흔히 행하는 잘못된 경향은 통계검정을 여러 개 실행한 후 그중 가장 유의미한 결과만 발표하고 마치 그것이 유일하게 행해진 검정인 것처럼 해석하는 것이다.**

  - 바이오 벤처회사 인터뮨의 CEO였던 스콧 하코넨은 임상시험에서 신약이 전반적인 효능을 보여주지 못했지만, 증상이 약한 일부 그룹에 한해서 신약의 효과가 유의미하게 있는 것으로 나타나자 투자자들에게 선별적으로 유의미한 결과만 발표하였다.

  - 후속 임상시험은 이 그룹에서조차 신약의 효능을 확인하지 못했고 이후 하코넨은 사기죄로 유죄판결을 받게 된다.

오늘의 강의 요점

- 재현성의 위기
- 통계분석이 잘못되는 경우

재현성 위기와 연구윤리

2. P-Hacking

P-Hacking

- 2012년에 2,155명의 심리학자를 대상으로 한 무기명 설문조사에 따르면 이 중 2%는 데이터 조작을 한 적이 있다고 한다.
- 이러한 데이터 조작을 통하여 (통계적으로)유의미한 결과를 제공하는 것을 p-hacking, 혹은 data snooping이라고 한다.
- 펜실베니아 심리학과 교수인 우리 사이먼슨의 주요 연구 관심분야 중 하나가 이런 데이터 조작을 발견하는 것이다.
- 예를 들면 그는 한 연구에서 15명의 피실험자로 구성된 3개의 다른 그룹에서 나온 표준편차가 모두 25.11로 같다는 사실에 주목하고 실제 자료를 바탕으로 컴퓨터 시뮬레이션을 통해 이런 일이 발생할 가능성이 거의 없음을 보였다. 그 연구책임자는 결국 사임하였다.

연구 부정행위

- 데이터를 조작한 것이 아니더라도 많은 연구들이 다양한 분야의 전공자들의 협업으로 이루어지고 있으며 실제 연구실험설계에서 분석과 최종결론이 이르는 단계를 공동연구자조차도 알지 못 하는 일이 비일비재하다.
- 뿐만 아니라 데이터 분석과정에는 수많은 의사결정(예를 들면 연속형 변수인 나이를 범주형으로 바꿀 때 나이의 범위를 정하는 경우와 이상점을 제외할 지 여부를 결정하는 경우)을 해야 한다. 사이먼슨은 이런 결정을 연구자의 자유도라는 표현을 사용했다.

연구 부정행위

- 이런 결정을 어떤 단계에서 자유롭게 할 수 있는지에 대한 답변은 지금 하고있는 분석이 탐색적 분석(Exploratory Data Analysis)인지 혹은 확증적 분석(Confirmatory Data Analysis)인지 여부에 달렸다. 탐색적 분석에서는 연구자의 자유도가 존중되지만 확증적 분석에서는 사전계획에 따라 분석이 진행되어야 한다.

비틀즈의 노래 When I’m Sixty-Four를 들으면 젊어지는가?

- 사이먼슨은 P-Hacking이 어떻게 이루어지는지는 보여주기 위해 다음과 같은 실험을 하였다.
- 먼저 펜실베니아 대학교 재학생들에게 비틀즈가 부른 “When I’m Sixty-Four”, “Kalimba”와 위글스가 부른 “Hot potato” 중 하나를 무작위로 들려준 후 수많은 의미 없는 질문(예를 들며 외식은 얼마나 즐기나, 100의 제곱근은 얼마인가, 아버지의 나이는 얼마인가 등)을 던졌다.

비틀즈의 노래 When I’m Sixty-Four를 들으면 젊어지는가?

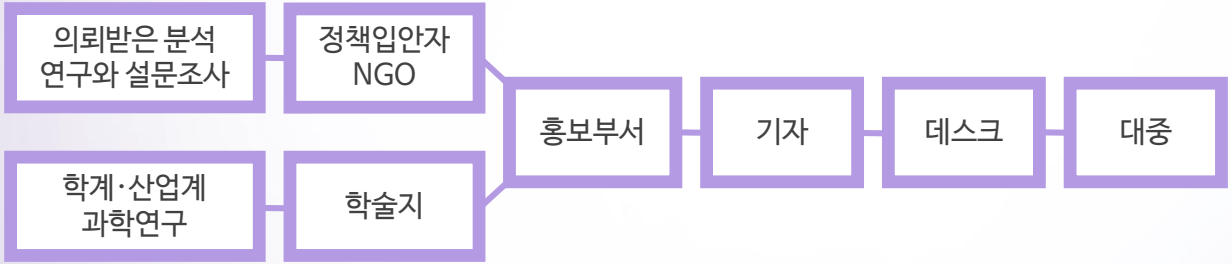
- 이후에 설문자료를 가능한 모든 방법을 동원하여 분석하고 통계적 유의성을 발견할 때까지 계속 피실험자를 추가하였다. 피실험자가 34명을 넘어서자 참가자들의 나이를 반응변수로 “When I’m Sixty-Four”와 “Kalimba”(Hot potato 제외)의 선택을 예측변수로 한 후 아버지의 나이를 통제한 회귀분석을 실시한 결과 회귀직선의 기울기가 통계적으로 유의미하다는 결론을 도출했다!

연구 부정행위는 얼마나 자주 일어나는가?

- **앞의 심리학자 대상 설문조사에서는 2%만이 데이터를 조작한 적이 있다고 시인했다.**
- **하지만 보다 광범위한 내용의 연구 부정행위에 대해서 물어본 결과 훨씬 많은 연구자들이 실제 연구 부정행위로 간주될 수 있는 행위에 참가했다는 것을 시인했다.**
  - 39%는 기대하지 않았던 결과를 처음부터 예측했던 것처럼 발표했다.
  - 58%는 결과가 유의미하지 않을 경우 데이터를 추가했다.
  - 67%는 연구의 모든 결과를 발표하지 않았다.
  - 94%는 이런 다양한 연구 부정행위 중 적어도 하나는 한 적이 있다고 시인했다.

통계분석 결과물은 어떤 방식으로 발표되어야 하나?

- 아래 그림은 통계분석 결과물이 어떤 방식으로 일반 대중에게 전달되는지를 보여주는 그림이다. 이러한 여러 단계를 거치면서 결과물의 해석에 대한 오류와 왜곡은 어디서나 생길 수 있다.



[통계분석 결과물이 일반 대중에 전달되는 과정]
(The Art of Statistics, p354)

기타 출처 #1

발견편향

- 앞의 전달과정에서 첫 번째 관문은 통계분석 결과의 출판에서 시작된다.
- 결과가 흥미롭지 않거나 연구수행기관(예를 들면 제약회사)의 목적에 부합되지 않는다는 이유로 발표되지 않는 연구 결과가 많다.
- 이로 인해서 통계적으로 유의미한 결과만 출판이 되는 발견편향(positive bias)이 존재한다.
- Journal of Negative Findings in Biomedicine과 같이 통계적으로 유의미하지 않은 결과만을 출판하는 학술지도 등장하였다.

홍보는 어떻게 해야 하나?

- **연구 결과에 관해 의욕 넘치는 보도자료 때문에 “왜 대학에 가면 뇌종양에 걸릴 위험이 커지는가?”와 같은 기사를 낳게 되는 참사가 벌어진다.**
- **2011년 영국 대학에 배포한 462개의 보도자료를 조사한 연구 결과는 다음과 같았다.**

1. 40%가 충고를 과장했다.
2. 33%가 인과관계를 과장했다.
3. 36%가 동물 연구 결과를 인간의 경우로 확대해석했다.
4. 언론의 과장된 표현 대부분은 보도자료에서 나온 것이다.

미디어의 보도방식

- 과학 기사나 통계분석을 다루는 기사 내용이 부실하다고 기자를 비난하는 경우 종종 있다.
- 특히 클릭 수를 유발하는 기사를 작성하기 위해 자극적인 기사 제목에 다는 경우에는 그 비난이 심해지기도 한다. 하지만 많은 경우 기사 제목은 기자보다 데스크에서 결정되는 경우가 많다.
- 언론 보도의 가장 큰 문제는 부적절한 결과 해석에 따라 사실 왜곡이 이루어진다는 점이다. 다음 목록은 미디어에서 행해지는 다양한 왜곡 보도 방식이다.

1. 현재 합의에 반하는 이야기를 하라.
2. 연구의 질에 구애 받지 않고 이야기를 홍보하라.

미디어의 보도방식

3. 불확실성을 발표하지 말라.
4. 장기간의 동향같은 전후사정이나 비교를 통한 관점을 제공하지 말라.
5. 단지 하나의 연관성이 관측될 때 원인을 제안하라.
6. 결과들의 관련성과 중요성을 과장하라.
7. 증거가 특정 정책을 뒷받침한다고 주장하라.

미디어의 보도방식

8. 목표가 안심시키는 것인가 아니면 겁주는 것인가에 따라 긍정적 또는 부정적 프레이밍을 사용하라.
9. 상충되는 관심이나 다른 기각은 무시하라.
10. 생생하지만 정보는 주지 않은 시각화를 활용하라.
11. 상대위험도만 제공하고 절대위험도는 제공하지 말라.

아무 생각 없이 TV시청하다가 죽을 수 있다?

- 한 연구에 의하면 하룻밤에 2시간 반 이하로 TV 시청을 하는 사람과 비교해서 5시간 이상 TV를 시청하는 사람들이 폐색전에 걸릴 상대위험도가 2.5라고 추정했다.
- 하지만 고위험군에서 조차 폐색전의 절대위험도는 매년 15만 8,000명 중 13명꼴이었다.
- 다시 말하자면 1만 2천 년 동안 매일 밤 TV를 5시간 이상 봤을 때 아마도 폐색전이 걸릴 것이라고 얘기할 수 있다.
- 절대위험도를 이야기하지 않고 상대위험도만 이야기하는 것은 자극적인 기사/홍보 제목은 달 수 있지만 많은 경우 진실과는 차이가 있다.

다시 대릴 벰의 연구로…

- 대릴 벰은 행한 많은 연구 중 재현 가능한 연구는 없었다.
- 그렇다면 대릴 벰은 어떻게 그런 연구결과를 얻을 수 있었을까?
- 벰은 수집한 데이터에 맞추어서 실험 설계를 나중에 변경하거나 여러 가지 가설 중 통계적으로 유의미한 결과만 보고하였다. 즉 여러 가지 이미지를 보여주고 그중 에로틱한 이미지에서만 통계적으로 유의미한 결과가 나오자 그 것만 보고하는 방식이었다.
- 아이러니하게 그의 2011년 연구를 통해 과학계가 재현 연구의 필요성을 자각하는 계기가 되었다.

오늘의 강의 요점

- 연구 부정행위
- 커뮤니케이션

출처

#1 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

재현성 위기와 연구윤리

3. 통계학으로 대화하기

난소암 혈청검사는 효과가 있는가?

- 2015년 영국에서 난소암에 관한 새로운 혈청검사에 관한 임상시험 결과가 의학 저널 Lancet에 발표되었다.
- 이 임상시험은 2001년부터 시작했으면 20만 명의 여성이 실험군과 대조군에 임의로 배정되었다.
- 연구팀은 추적 조사 기간 동안 검사를 통해서 난소암 사망률이 낮아질 것으로 생각했으며 Cox's proportional hazards model을 사용하여 검사의 사망률 감소 효과를 추정하였다.
- 연구결과 실험군과 대조군의 사망률차이는 있지만, 통계적으로 유의미하지 않다는 결론을 제시했다. 하지만 영국의 주요 일간지인 인디펜던트는 "새로운 검사 방법의 큰 성공이 영국의 국가 암 검진사업으로 이어질 것이다"라는 기사를 내었다!

데이터 문해력

- **인디펜던트지가 이런 기사를 낸 이유는 무엇일까?**
- **이 문제에 대한 해답을 찾기 위해 이러한 연구 결과를 일반 대중에게 올바르게 전달하기 위해 어떤 노력을 해야 하는지 알아보자.**
- **이러한 노력은 아래 3그룹의 협조하에 가능하다.**
  - 통계분석자: 과학자, 통계학자, 설문조사 회사 등
  - 전달자: 과학 저널, 정부 기관, 홍보부처, 기자
  - 독자: 정책입안자, 전문가 그룹 등 통계분석 결과를 이용하고자 하는 자

통계 분석자의 역할: 난소암 혈청검사는 효과가 있는가?

- 통계분석자의 역할은 사전에 준비된 분석계획에 따른 결과를 발표해야 한다는 점이다. 하지만 많은 경우 논문이나 언론에 발표된 결과는 사전분석계획보다 분석내용이 관측된 데이터에 맞춰서 조정되는 경우가 많다.
- 난소암 혈청검사 임상시험에서 연구팀은 모든 임상시험참가자를 대상으로 실험군과 대조군의 사망률 차이를 비교했는데 일부 참가자들이 임상시험 시작하기 전에 난소암을 가지고 있는 것이 발견되었다.

통계 분석자의 역할: 난소암 혈청검사는 효과가 있는가?

- 이런 난소암 환자를 분석대상에서 제외한 결과 실험군에서 난소암 사망률이 20% 감소하였다(p-value =0.02).
- 또한 모든 참가자들을 대상을 할 때에도 무작위 배정 이후 7년이 지나면 사망률이 23% 감소했다는 사실을 발견했다.

통계 분석자의 역할: 난소암 혈청검사는 효과가 있는가?

- 임상시험 분석계획을 만들 때에는 참가자 중에 난소암환자가 있을 수 있다는 것을 생각하지 못해서 참여범위를 정하는 기준에 포함시키지 않았으며 검사결과의 효과가 어느정도 시간이 걸릴 수 있다는 것을 짐작하지 못해서 기간을 나누어서 분석하는 것을 고려하지 않았다.
- 하지만 실제 데이터를 받았을 때 원래 분석계획이 미진하다고 생각하여 추가분석을 실시하였으며 연구결과 발표 시 원래 계획에 충실한 분석발표를 하면서 추가발표로 난소암 환자를 제거한 경우와 기간별 분석을 포함하였다.

통계 분석자의 역할: 난소암 혈청검사는 효과가 있는가?

- 대부분의 언론은 원래 분석계획에 따른 결과만 보도하였지만 인디펜던트지만 유일하게 연구의 결론을 잘 반영하였다.
- 여기서 통계분석가의 역할은 분석계획이 미진할 경우 원래 계획에 따른 분석결과와 함께 추가분석의 내용을 같이 포함해야 한다는 점이다.

전달자의 역할: 데이터 저널리즘

- 2019년 뉴욕타임즈는 데이터 에디터자리를 신설하고 워싱턴 대학 통계학과 박사과정 중퇴생 출신의 아만다 콕스를 그 자리에 임명한다.
- 데이터 에디터는 데이터 시각화와 뉴욕타임즈의 업샷 세션(여러 가지 이슈에 대해 데이터 시각화와 다양한 통계모형을 통한 분석내용을 올리는 웹사이트)을 관장하는 일을 한다.
- 지나치게 단순화된 파이 차트와 너무 현란한 데이터 시각화의 사이에서 적당한 조화를 통해 독자들이 이해하기 쉽게 정보를 전달하는 것이 중요하다.
- 독감백신 유해론과 같은 기사에서 나오는 낮은 수준의 통계 관련 논쟁은 한국 데이터 저널리즘의 현주소를 보여주고 있다.

매력적인 사람일수록 딸을 많이 낳는다?

- 2007년 영국 런던정경대 카나자와 교수는 Journal of Theoretical Biology에 매력적인 부모일수록 딸을 많이 낳는다는 논문을 게재하였다. 그는 사실 2005년과 2006년에 같은 저널에 키가 큰 부모가 아들을 많이 낳는다는 논문과 폭력적 성향의 남자가 아들을 가질 경우가 더 많다는 논문을 게재하였다.

매력적인 사람일수록 딸을 많이 낳는다?

- 이 가설을 검증하기 위해 그는 미국의 한 설문조사에서 본인의 외모를 1점부터 5점까지 평가한 한 사람들을 대상으로 15년 후 그 사람들의 첫 번째 자녀의 성별을 조사하였다. 그 결과 외모를 최상으로 평가한 사람들은 첫 번째 자녀의 52%가 딸이었고 나머지 등급의 사람들의 경우 첫 번째 자녀가 딸인 비율은 44%였다.
- 이 차이는 통계적으로 유의하다. 참고로 장원철 교수의 첫 번째 자녀는 아들이다.

매력적인 사람일수록 딸을 많이 낳는다?

- 이 논문의 문제점에 관해서 컬럼비아 대학의 앤드류 겔만 교수는 “Of Beauty, sex, and power: Statistical challen-ges in small effects”라는 논문에서 조목조목 지적했다.
- 가장 큰 문제점은 분석 방법에 있다. 이 경우는 반응변수를 성별로 하고 예측변수를 외모 점수로(로지스틱) 회귀분석을 실시하는 것이 보다 적합하다.

매력적인 사람일수록 딸을 많이 낳는다?

- 겔만 교수는 추가로 피플지가 매년 발표하는 세계에서 가장 아름다운 50인의 5년 치 자료를 이용하여 이 사람들의 경우 첫 번째 자녀의 여성 비율이 47.7%라는 점을 제시한다.
- 이 사례는 권위 있는 저널에 출간된 결과라도 하더라도 맹목적으로 믿을 수 없다는 사실과 너무 자극적인 결과를 발표하는 논문이라면 조금 더 자세히 분석과정을 살펴보아야 한다는 점이다.

출판편향과 p값 분포

- 적절하지 않은 분석 혹은 연구 부정행위가 개입이 된 경우 출판 편향이 존재할 수 있다.
- 예를 들면 저널에서는 주로 새로운 발견에만 관심이 있다. 즉 귀무가설을 기각하는 연구일수록 논문이 게재될 가능성이 많고 그 결과 많은 연구자들이 p값이 0.05보다 작은지 여부에만 관심이 집중된다.

출판편향과 p값 분포

- p값의 분포를 통해서 출판편향이 존재하는지 여부를 알아볼 수 있다. 만약 귀무가설이 맞다면 p값의 분포는 0과 1사이의 균등분포를 따른다는 것이 알려져 있다. 따라서 어떤 논문에서 여러 가지 연구결과를 보고할 때가 연구결과들이 모두 실질적으로 효과가 없는 경우라면 관련된 연구의 p값은 균등분포를 따라야 한다. 반대로 효과가 있는 경우가 많다면 관련 연구의 p값은 작은 값들에 치우친 경향을 보일 것이다.

출판편향과 p값 분포

- **출판편향을 알아보기 위해 p값의 분포에서 다음 2가지 사항을 살펴보자. 실제 출판된 대부분의 논문에서 p값은 0.05보다 작기 때문에 우리는 0.05 이하의 p값의 분포를 살펴본다.**

  - 0.05보다 살짝 작은 p값들이 무더기로 있다면 유의수준보다 작게 p값을 만들기 위해 일부 실험이 조작되었을 가능성도 있다.

  - p값이 0과 0.05 사이에 균등하게 있다면 이것은 제1종의 오류의 결과로 우연히 귀무가설을 기각한 것으로 간주해야 한다.

데이터 윤리

- 사람에게 영향을 미치는 알고리즘은 공정하고 투명해야 한다.
- 과학연구는 정직하고 재현 가능해야 한다.
- 통계자료의 전달은 신뢰할 수 있어야 한다.
- 사생활 보호와 데이터 소유권도 생각해 보아야 할 문제들이다

오늘의 강의 요점

- **데이터 문해력**
  - 통계분석가
  - 전달자
  - 독자
- **출판편향과 p값**
- **데이터 윤리**

재현성 위기와 연구윤리

Lab 13. 통계분석을 잘 하려면?

통계학 기반 주장을 평가하는 방법

- **데이터 기반 주장은 다음 조건을 만족해야 한다.**
  - 접근 가능: 정보를 쉽게 얻을 수 있어야 한다.
  - 이해 가능: 정보를 쉽게 이해할 수 있어야 한다.
  - 평가 가능: 주장의 신빙성을 확인할 수 있어야 한다.
  - 사용 가능: 원한다면 정보를 활용하여 다른 목적으로 사용할 수 있어야 한다.

통계기반 주장을 점검하기 위한 10가지 질문

연구 결과에 대한 신뢰 여부

1. 관련 연구는 얼마나 엄밀히 수행되었는가? 내적타당성, 질문의 단어 선택, 표본의 대표성 등을 점검하자.
2. 결과에서 통계적 불확실성은 무엇인가? 오차범위, 신뢰구간, 통계적 유의성, 다중비교 등에 관해서 점검하자
3. 요약은 적절한가? 평균, 분산, 상대위험도, 절대위험도가 모두 제시되었는지 살펴보자.

통계기반 주장을 점검하기 위한 10가지 질문

출처는 얼마나 믿을 만 한가?

4. 이야기의 출처는 얼마나 믿을 만한가? 유튜브나 개인 블로그라면 이야기하지 않는 것이 낫다.

5. 이야기를 장황하게 늘어놓고 있는가? 본질을 흐리는 과장된 사례에 주목하지 말자.

6. 들려주지 않는 것은 무엇인가? 모든 이야기는 양측의 다른 관점이 있다. 들려주지 않은 이야기가 가장 중요한 이야기일 수 있다.

통계기반 주장을 점검하기 위한 10가지 질문

결과에 관한 해석은 얼마나 믿을 만 한가?

7. 그 주장이 알려진 것들과 얼마나 잘 들어맞는가? 기존의 다른 연구 결과와 비교하여 종합적으로 고려해보자

8. 보인 것에 대한 설명으로 무엇을 주장하는가? 상관관계와 인과관계, 평균으로의 회귀, 검사의 오류와 같은 실수를 하지 않는지 확인하자.

9. 들려주는 이야기와 청중과의 연관성은 무엇인가? 쥐에 관한 실험 결과를 사람에 관한 결과로 확장했을 경우 일반화가 가능한지 생각해보자.

10. 주장하고 있는 바의 영향력은 어느 정도 인가? 통계적 유의성과 실질적 유의성의 차이에 대해서 생각해보자.

통계분석을 잘하는 10가지 규칙

- **아래 10가지 규칙은 다음 논문의 요약이다. Ten Simple Rules for Effective Statistical Practice**

1. 통계분석 방법은 데이터로 주어진 과학적 질문에 답변을 제공하게 해야 한다.
2. 신호는 항상 소음과 같이 나타난다.
3. 항상 미리 상세한 연구계획을 준비하라.
4. 데이터의 질에 신경을 써라
5. 통계분석은 단순히 통계 패키지를 돌리는 것이 아니다.

기타 출처 #1

통계분석을 잘하는 10가지 규칙

6. 단순한 모형이 좋다.
7. 불확실성에 대한 근거를 제시하라.
8. 가정을 항상 확인하라.
9. 새로운 데이터를 사용해서 분석 절차를 반복해라
10. 재현 가능한 연구가 될 수 있도록 분석 결과를 제공하라.

기타 출처 #1

재현 가능한 연구를 위한 수단

- 프로그램에 충분히 주석을 달아라 - 프로그래머/분석가의 가장 큰 적은 한 달 전에 주석을 달아놓지 않은 자신이다.
- 버전 관리를 하라 - 프로그램/문서의 변천사를 기록하여 특정 시점의 프로그램/문서를 나중에 쉽게 찾아볼 수 있게 관리하는 것이 중요하다.
- 버전 관리는 git/github의 사용을 통해서 이루어진다.
- R/RStudio 사용 시 R Markdown을 사용하여 분석과 보고서 작성이 동시에 이루어 지도록 하라.

오늘의 강의 요점

- 통계학 기반 주장을 평가하는 방법
- 통계분석을 잘하는 10가지 규칙
- 재현 가능한 연구
  - Markdown
  - Version Control과 Github

출처

#1 Kass et al. (2016). Ten Simple Rules for Effective Statistical Practice, PLoS Computational Biology 12:6, e1004961