데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

기타 출처 #1

통계학의 소개와 자료수집

0. 들어가면서

기타 출처 #1

누구를 위해서 통계학을 배우나?

전통적인 통계학 과목 개요

- 자료의 소개와 요약 통계량
- 확률과 확률분포
- 표본분포
- 각종 검정 방법 (t, Z, χ^2, F)
- 실제 사용 사례

기타 출처 #1

누구를 위해서 통계학을 배우나?

데이터 사이언스를 위한 통계학

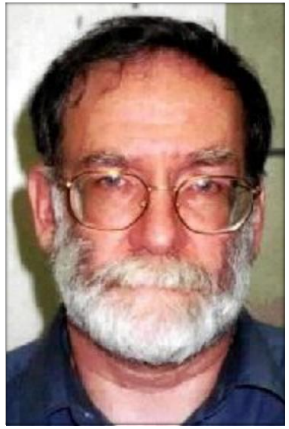
- 실제 문제를 통한 동기유발
- 데이터 시각화와 탐색적 자료 분석
- 데이터를 통해서 알아낼 수 있는 지식에 집중

  (내재되어 있는 편향, 인과관계 등)
- 모형과 알고리즘
- 확률 이론 기반의(예측에 관한) 불확실성의 근거 제시

기타 출처 #2

통계학을 알면 연쇄살인을 막을 수 있다!

헤롤드 시프먼을 아시나요?



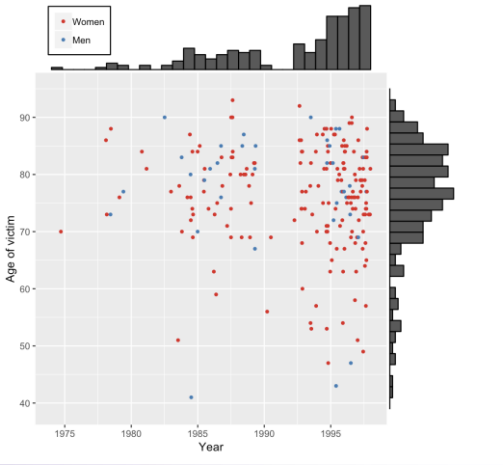
[헤롤드 시프먼]
[Harold Shipman]

- 헤롤드 시프먼을 영국 역사상 가장 많은 사람을 살해한 연쇄살인범으로 맨체스터 교외에서 온화한 가정의(family doctor)로 알려져 있었다.
- 1975년에서 1998년 사이 적어도 215명을 살해한 것으로 믿어지며 45명의 추가 살인도 의심되었다.
- 희생자 중 한 명의 유언장을 위조한 것을 희생자의 딸이 발견하여 경찰에 신고하여 덜미를 잡히게 된다.

기타 출처 #3

통계학을 알면 연쇄살인을 막을 수 있다!

헤롤드 시프먼의 희생자들의 특징



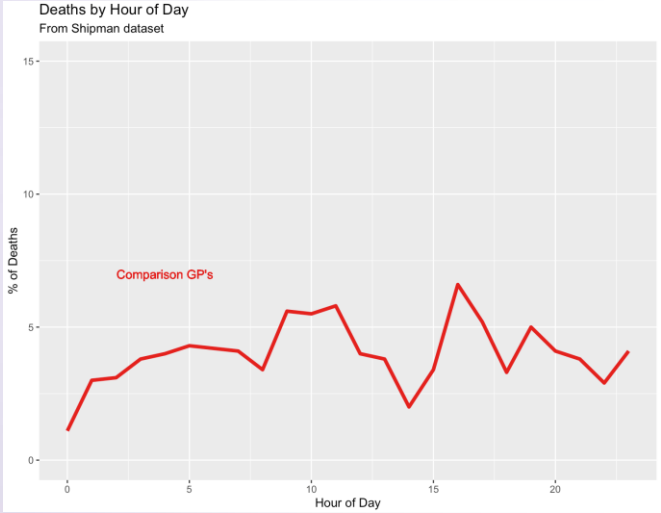
[헤롤드 시프먼에 의해서 살해된 215명의 나이와 사망연도를 보여주는 산점도와 막대그래프]
(Statistics, 2019, p 3)

- 왼쪽 그림은 헤롤드 시프먼에 의해 살해된 215명 희생자의 성별/연령별 분포를 보여준다.
- 여성이 남성보다 많다는 것을 알 있으면 희생자의 연령의 대부분은 70-80대였다.
- 1992년 이전에는 시프먼이 공동진료를 하던 시기였고 이후 단독 개원을 한 후 범행 건수가 폭발적으로 증가한 것을 알 수 있다.

기타 출처 #4

통계학을 알면 연쇄살인을 막을 수 있다!

헤롤드 시프먼의 희생자들은 언제 살해되었나?



[같은 지역 일반 가정의 환자들의 시간대별 사망비율]
(Statistics, 2019, p 3)

기타 출처 #5

통계학을 알면 연쇄살인을 막을 수 있다!

헤롤드 시프먼의 희생자들은 언제 살해되었나?

[같은 지역 일반 가정의와 시프먼 환자들의 시간대별 사망비율]
(Statistics, 2019, p 5)

기타 출처 #6

통계학을 알면 연쇄살인을 막을 수 있다!

헤롤드 시프먼을 (데이터 기반 증거로) 좀 더 일찍 체포할 수 있었을까?

- 시프먼의 환자 사망자가 다른 가정의와 비교해서 심각하게 많아진 시점을 알 수 있다면 우리는 시프먼의 진료행위가 수상하다는 합리적 의심을 할 수 있다.
- 실제 SPRT(Sequential Probability Ratio Test)라는 방법을 이용해서 이 시점을 찾을 수 있다.
- SPRT는 통계적 품질관리에서 생산라인의 관리를 위해서 많이 사용되는 기법이다.

통계학의 역할

헤롤드 시프먼을 (데이터 기반 증거로) 좀 더 일찍 체포할 수 있었을까?

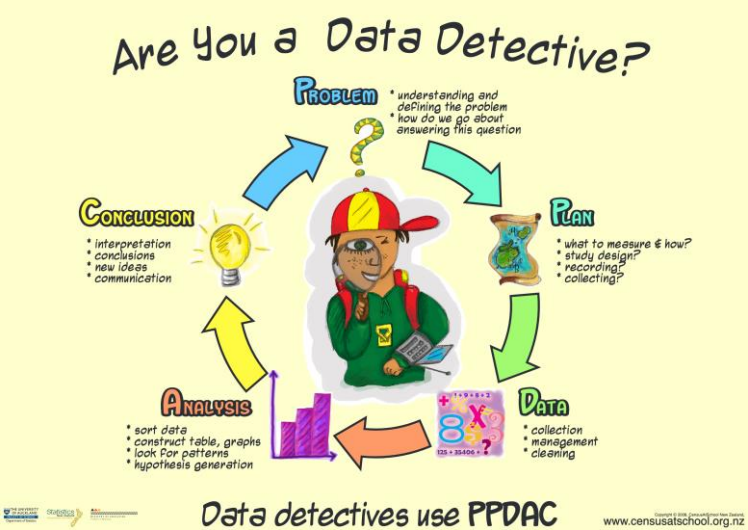
- **데이터의 불완전성**
  - ↳ 측정 도구의 불확실성

    : 응급실에서 아픈 정도를 1에서 10까지 물어볼 경우 대답은?
  - ↳ 데이터에 내재된 변동성

    : 신호+소음으로 이루어진 데이터에서 소음으로 인한 변동성
- **통계분석이란?**
- **재현 가능성의 위기**
- **데이터 문해력 (Data Literacy)**

데이터 탐정

Problem-Plan-Data-Analysis-Conclusion



[Data Detective Poster]

사진 출처 #7

PPDAC의 실제 적용

헤롤드 시프먼의 사례연구의 중심으로

- Problem: 헤롤드 시프먼은 얼마나 많은 사람을 살인했는가?
- Plan: 시프먼 환자 중 사망한 사람들의 특징(사망 시간, 나이 등)에 나타내는 자료를 수집
- Data: 시프먼 환자 자료 중 cleaning이 필요한 자료가 있는지 확인
- Analysis: 데이터 시각화(탐색적 자료 분석)와 가설검정(확증적 자료 분석)
- Conclusion: 시프먼은 최소 215명을 살해했으며 45명이 추가로 살해되었을 가능성이 있다.

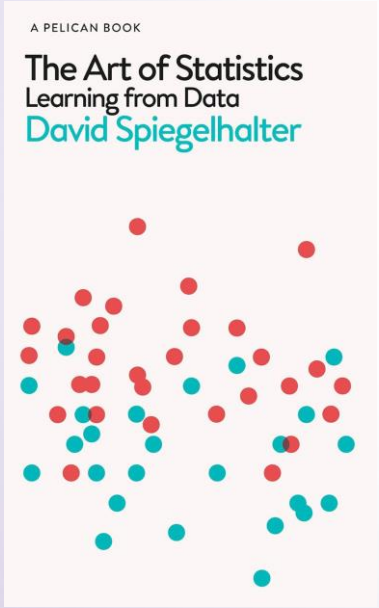
이 과목을 통해서 알고자 하는 것은?

앞으로 다룰 문제들

- 힉스 입자는 존재하는가?
- 더 많은 사람이 찾는 병원이 과연 치료를 잘하는가?
- 스타틴은 복용 효과가 정말 있는가? 있다면 얼마나 효과가 있는가?
- 타이타닉에서 가장 운이 좋은 생존자는 누구인가?
- 리차드 3세의 해골은 정말 발견된 건인가?

강의 교재

The Art of Statistics/숫자에 약한 사람들을 위한 통계학 수업



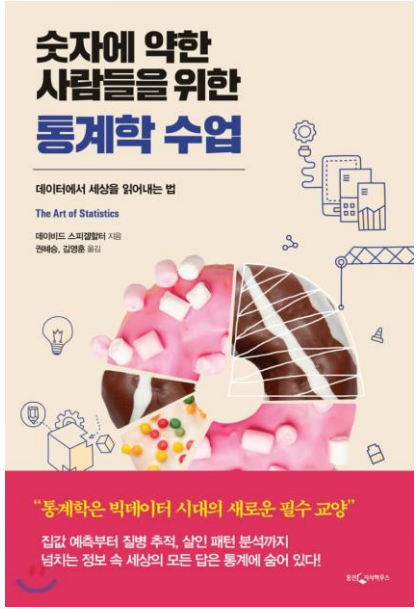


사진 출처 #8, 9

강의 교재

OpenIntro Statistics

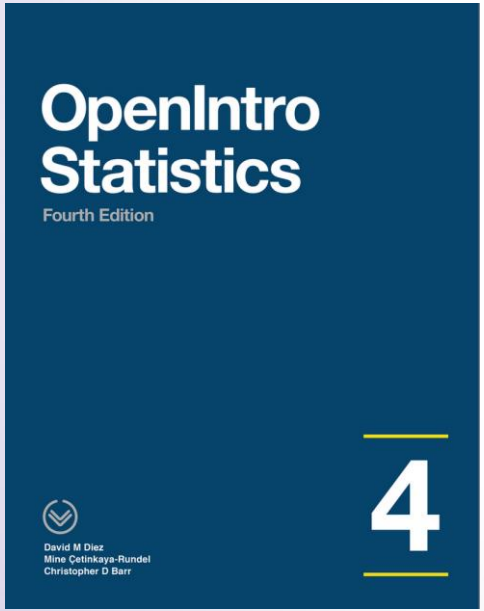


사진 출처 #10

오늘의 강의 요점

- Problem-Plan-Data-Analysis-Conclusion 접근방식
- 데이터 문해력의 중요성
- 강의에 사용된 그림은 The Art of Statistics의 github페이지 (https://github.com/dspiegel29/ArtofStatistics)에 제공되는 R code를 사용하여 생성되었다.

출처

#1~2 데이비드 스피겔헬터 교수의 LSE 강연(https://bit.ly/2RLsCOI)중에서

#3 위키피디아 https://bit.ly/2ZVLWg

#4~6 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House,

#7 https://new.censusatschool.org.nz/resource/data-detective-poster/

#8 https://amzn.to/343rk7J

#9 https://bit.ly/348BmEm

#10 https://leanpub.com/openintro-statistics

통계학의 소개와 자료수집

Lab 1: PPDAC 연습

현재 국내 실업률은 얼마인가?

Problem 정의

- 실업률이란 무엇인가?
- 영국 기준: 1979년부터 1996년 사이 31번 이상 바뀜
- 한국 기준: 경제활동인구(취업자+실업자)에서 실업자가 차지하는 비율. 여기서 경제활동인구란 만 15세 이상 인구 중 상품이나 서비스를 생산하기 위하여 실제로 수입이 있는 일을 한 취업자와 일을 하지 않았으나 그 일을 즉시 하려고 구직활동을 하는 실업자를 합하여 통칭하는 용어이다.

지구상에는 얼마나 많은 나무가 있을까?

PPDAC: Problem 단계

- 나무란 무엇인가?
- 나무의 정의: 사람 가슴 높이에서 잰 나무줄기의 지름이 충분히 크고 딱딱한 줄기를 가진 식물
- 여기서 나무줄기 지름의 큰지 여부의 기준은 대부분의 나라에서는 10cm

지구상에는 얼마나 많은 나무가 있을까?

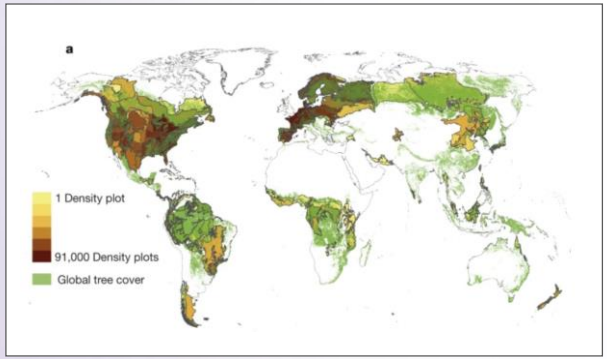
PPDAC: Plan 단계

- **어떻게 측정할 것인가?**
  - 딱딱한 줄기를 가진 식물을 하나하나 측정하는 것은 불가능!
- **다음과 같은 방법으로 자료를 구한다.**
  - 지리적, 기후적으로 유사한 일련의 지역들(biome)별로 나무의 개수를 센 후 지역별로 단위면적당 나무 숫자의 평균을 구하고 지역별 GIS 관련 변수들의 정보도 수집한다.
  - 위성사진을 이용하여 각 유형별 지역(biome)이 지구 전체에서 차지하고 있는 면적을 추정

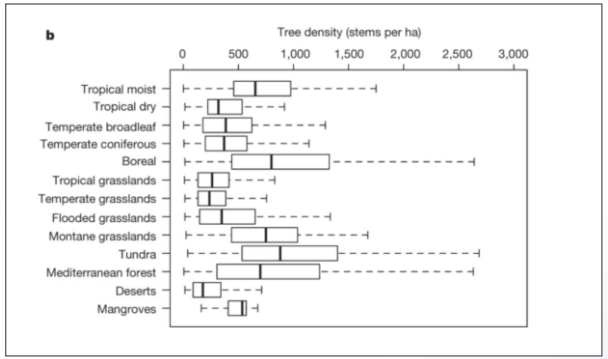
지구상에는 얼마나 많은 나무가 있을까?

PPDAC: Data 단계

- 다양한 international forestry database를 사용하여 약 43만 군데의 tree density 측정 결과와 관련된 GIS 변수 수집



[Biome level forest density]



[각 유형별 단위면적당 나무 수]

[Map of data points and raw biome-level forest density data]

기타 출처 #1

지구상에는 얼마나 많은 나무가 있을까?

PPDAC: Analysis 단계

- 최종 수집한 자료를 바탕으로 통계모형(음이항 회귀분석)을 이용하여 전체 나무의 수 추정

기타 출처 #2

지구상에는 얼마나 많은 나무가 있을까?

PPDAC: Conclusion 단계

- 전체 나무의 숫자는 약 3조 400억 ± 1,000억
- 매년 150억 그루의 나무가 지구상에서 잘려 나가는 것으로 추정
- 인류문명이 시작한 이래 지구상 나무의 46%가 사라졌다고 추정

기타 출처 #3

출처

#1~3 T. W. Crowther, (2015) Mapping tree density at a global scale, Nature, PowerPoint slide for Fig.1 (https://go.nature.com/3kAufLz)

통계학의 소개와 자료수집

1. 데이터는 어떻게 모아지는가? I

1.1 실험

Challenger: The Final Flight

O-ring 의 저주

[Challenger : The Final Flight]

- 1986 년 1월 28일 우주왕복선 챌린저호가 발사 후 73초 만에 고체연료 추진기 고장으로 인한 폭발로 7명의 승무원 전원이 사망한다.
- 사고의 원인은 보조 추진 로켓의 O 링(고무 패킹)이 추운 날씨로 인해 얼어버려 제 기능을 하지 못했기 때문으로 밝혀졌다.

사진 출처 #1

Challenger: The Final Flight(리차드 파인만의 실험)

- 사고원인을 규명하기 위해 열린 상원 청문회에서 리차드 파인만은 간단한 실험을 통해서 O-ring이 낮은 온도에서 탄력성을 잃게 되는 것을 보인다.
- 이 실험의 문제점은 무엇인가?

[O-ring]

[파인만교수의 상원청문회에서 O-ring 실험]
(Visual explanations, 2005, p. 51)

사진 출처 #2, 기타 출처#3

실험이란?

- **실험은 우리가 의미 있는 결론을 도출하기 위한 수단이다.**
- **다음과 같은 과학 연구 가설(Problem)을 생각해보자.**
  - 추운 날씨가 우주왕복선의 O-ring의 탄력성을 잃게 한 원인인가?
  - 스텐트 시술이 뇌졸중을 예방하는 데 효과적인가?
  - 중국산 한약재와 국산 한약재를 어떻게 구별할 수 있나?
  - 새로 개발한 코로나 -19 백신이 실제로 효용이 있는가?

연구계획(Study Design)

- **파인만의 실험에서 추가로 필요했던 것은 얼음을 넣지 않은 컵에서 고무밴드가 탄력성을 유지하는지 여부에 대한 확인이다.**
- **일반적으로 연구설계에서 실험대상은 실험군(treatment group) 혹은 대조군(control group)중 하나에 속하게 된다.**
  - 실험군(Treatment group): 고무밴드를 얼음물(treatment)에 집어넣는다 .
  - 대조군(Control group): 고무밴드를 미지근한 물에 집어넣는다 .
  - 반응변수(Response variable): 고무밴드의 탄력성
  - 설명변수(Explanatory variable): 물의 온도

설명변수와 반응변수

- 한 쌍의 변수(variable)가 주어진 경우 한 변수가 다른 변수에 영향을 주는 경우 전자를 설명변수, 후자를 반응 변수라고 한다 .
- 두 변수 사이의 연관성을 보이는 것이 반드시 인과성을 보이는 것은 아니다(예 : 안전벨트와 비행기 진동).

소아마비 백신 임상시험

- 1954 년 미국의 National Foundation for Infantile Paralysis(NFIP)는 그 전해 피츠버그 대학의 조나크 소크 박사가 개발한 소아마비 백신을 시험하려고 한다.
- NFIP 는 먼저 몇 개의 학군을 선정한 후 각 학군에서 1, 2, 3 학년 학생들을 대상으로 다음과 같은 임상시험을 실시하였다.

  ⤷ 학부모의 동의를 받은 2학년생 모두에게 백신을 투약했다.

  ⤷ 1학년과 3학년은 control group으로 사용하였다.

소아마비 백신 임상시험

- **백신 투약 결과 소아마비 발병률은 다음과 같았다.**
  - Treatment group(부모 동의를 받은 2학년): 25%
  - Control group(1학년과 3학년): 54%
  - 부모의 동의를 받지 않은 2학년: 44%
- **이 실험의 문제점은 무엇인가?**
- **부모의 동의 여부와 발병률과 관련이 있을 수 있다!**

소아마비 백신 임상시험

- 첫 번째 임상시험의 문제점을 깨달은 NFIP는 2차 임상시험을 단행하였다.
- 부모 동의를 받은 학생들 중 임의로 control group을 할당하였다.
- Control group으로 할당된 학생들에게도 위약(placebo)을 주고 담당 의사들에게도 학생들이 어떤 그룹에 속했는지 알려주지 않았다.

소아마비 백신 임상시험

- **2 차 임상시험의 결과는 다음과 같다.**
  - Treatment group: 28%
  - Control group: 71%
  - 부모의 동의를 받지 않은 2학년: 46%

실험계획의 원칙

효율적인 실험을 위해서 다음과 같은 원칙을 지켜야 한다.

1. **Controlling** 관심이 있는 treatment를 받는 그룹과 control group을 비교한다. 임상시험에서 control group은 위약(placebo)를 받는다.

2. **임의할당(Randomization)** control group과 treatment group에 참가자를 임의로 할당한다.

3. **반복(Replication)** 충분히 큰 표본을 사용할 경우 설명 변수가 반응 변수에 미치는 영향에 대한 추정 결과를 반복적으로 관찰할 수 있다.

4. **이중 암맹(Double-blind)** 참가자와 연구자 모두 참가자가 속한 group이 어느 group인지 연구가 끝날 때까지 알려주지 않는다.

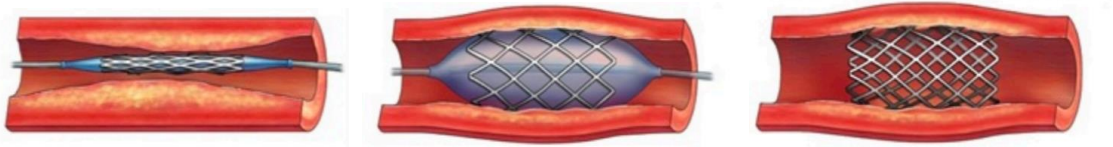
5. **블록화(Blocking)** 반응 변수에 영향을 미치는 다른 변수가 있을 경우 참가자들을 그 변숫값에 따라 block을 나눈 후 block 별로 참가자를 임의로 treatment group에 할당한다.

소아마비 백신 임상시험

- 소아마비 백신 1차 임상시험은 실험 계획의 어떤 원칙을 위배 했는가?
- 2차 임상시험에서 1학년과 3학년은 control group에 포함 하지 않은 이유는 무엇인가?

스텐트 시술은 뇌졸중을 예방하는가?

- 스텐트는 심장질환을 치료하기 위해 관상동맥 내부의 혈류를 유지하기 위해 혈관에 삽입하는 기구이다.
- 의사들은 스텐트 시술이 뇌졸중 예방에도 도움이 되는지 관심이 있다.



[스텐트 시술]

사진 출처 #4

대조군 vs 실험군

- **451명의 잠재적 뇌졸중 위험 환자를 대상으로 임의로 다음 두 그룹 중 하나에 배정하였다.**
  - 실험군: 환자들은 전통적인 의료관리와 스텐트 시술을 받는다.
  - 대조군: 환자들은 전통적인 의료관리만 받는다.
- **224명이 실험군에 227명이 대조군에 배정되었다.**

스텐트 시술은 뇌졸중을 예방하는가?

- 스텐트 시술의 효과를 측정하기 위해 실험을 시작한 후 30일 째와 1년 후 각 그룹에서 환자들의 뇌졸중 발병 여부를 기록하였다.

환자	그룹	0-30일	0-365일
1	실험군 (treatment)	no event	no event
⋮	⋮	⋮	⋮
451	대조군 (control)	no event	뇌졸중

[스텐트 임상시험결과]

사진 출처 #1

스텐트 시술은 뇌졸중을 예방하는가?

- 실험 결과는 다음과 같이 정리할 수 있다.

	0-30일		0-365일	
	뇌졸중	no event	뇌졸중	no event
실험군 (treatment)	33	191	45	179
대조군 (control)	13	214	28	199
합계	46	405	73	378

[스텐트 임상시험결과 정리]

기타 출처#5

스텐트 시술은 뇌졸중을 예방하는가?

- 이 실험 결과는 비율을 이용하여 요약할 수 있다.
- 실험군에서 1년 후 뇌졸중 환자의 비율은 45/224=0.2이다.
- 대조군에서 1년 후 뇌졸중 환자의 비율은 28/227=0.12이다.
- 스텐트가 뇌졸중을 예방하는 데 도움을 주었다고 할 수 있는가?

기타 출처#6

오늘의 강의 요점

실험 계획의 원칙

1. 실험군/대조군
2. 임의할당
3. 반복
4. 이중 암맹
5. 블록화

출처

#1 https://www.imdb.com/title/tt12930534/mediaviewer/rm4058751233

#2 https://en.wikipedia.org/wiki/O-ring

#3 E.R. Tufte, (2005), Visual explanations, Graphics Press, P.51

#4 https://bit.ly/3mJBUJi

#5~6 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel , M (2019) Open IntroStatistics, 4th edition, OpenIntro, Inc.p9

통계학의 소개와 자료수집

1. 데이터는 어떻게 모아지는가? II

1.2 관측연구

자료수집 유형

- **실험(Experiment)**

  연구자가 실험참가자를 임의로 다양한 조건하에 배치하여 설명변수와 반응변수 사이의 인과성(causality)을 조사한다.

- **관측연구(Obervational Study)**

  연구자가 자료를 관측하면서 수집하는 경우로 자료의 생성과정에 전혀 관여하지 않는다. 이 경우 설명변수와 반응변수의 연관성(association)을 밝히는데 초점을 둔다.

- **연관성이 인과성을 의미하는 것은 아니다!**

전향적 연구와 후향적 연구

전향적 연구(Prospective Study)

연구자가 대상자를 추적 관찰하면서 관련 정보를 얻는다. 예를 들면, 호흡기질환 혹은 암에 관한 연구를 하기 위하여 117,000명의 간호사를 대상으로 비만 정도를 기준으로 위험군과 일반군으로 나눈 후 두 그룹 간의 질병 발병률을 비교한다. 주로 실험과 관측연구 모두에서 사용된다.

후향적 연구(Retrospective Study)

이미 일어난 일에 대한 정보를 얻는 연구를 말한다. 예를 들면, 폐암 환자들을 대상으로 과거 흡연 여부를 알아본다. 주로 관측연구에서 사용된다.

흡연이 폐암을 유발하는가?

- 흡연이 폐암에 미치는 영향을 알고 싶다. 이 연구를 수행하기 위한 실험설계가 가능한가?
- 10세 아동 1,000명을 실험대상으로 모집한 후 randomization을 통해서 절반을 실험군, 나머지 절반을 대조군에 할당한 후 실험군에 속한 아동에게 흡연을 하게 한 후 10년간 추적 관찰하여 실험군과 대조군의 폐암 발병률을 비교한다.

흡연이 폐암을 유발하는가?

- 위의 실험은 윤리적으로 가능하지 않은 전향적 연구이므로 실제로 다음과 같은 후향적 연구를 진행한다. 폐암 환자 500명과 일반인 500명에게 흡연 여부를 물어본다. 이 경우 폐암 환자의 흡연율과 일반환자의 흡연율을 비교하게 할 수 있지만, 우리가 원하는 것은 흡연자 중 폐암 발병률과 비흡연자의 폐암 발병률의 비교이다!
- 위의 문제를 해결하기 위해 오즈비의 개념이 등장한다.

연관성(Association)과 인과성(Causation)

- **관측연구의 경우 연관성을 찾을 수 있다.**

  예: 폐암은 비흡연자보다 흡연자에게서 많이 발견된다. 따라서 폐암과 흡연은 강한 상관관계가 있다. 흡연과 폐암의 인과성은 1950년대 영국의 역학자 리처드 돌과 통계학자 오스틴 힐에 의해서 입증된다.

- **연관성은 인과성을 증명하기 위한 중간단계로 볼 수 있지만, 반드시 인과성을 의미하는 것은 아니다.**

  예: 비행기 좌석에 빨간 경고등이 켜지는 경우 비행기가 심하게 흔들린다. 하지만 경고등이 비행기를 흔들지는 않는다.

약을 꾸준히 복용하는 습관이 중요하다?

- 심근경색을 예방하는 약의 효능을 알아보기 위해 임상시험을 실시하였다. 대상은 심장질환이 있는 8,341명의 중년남성이었으며 randomization을 통해서 이 중 5,552명은 실험군에, 2,789명은 대조군에 할당되었다.
- 위약과 심근경색약(colfibrate)을 각각 실험군과 대조군에 속한 사람들에게 복용할 것을 권하고 5년 후에 사망률을 조사한 결과 실험군에서는 20%, 대조군에서는 21%의 사망률을 보였다.
- 약이 효능이 없는 이유 중 하나로 실험군과 대조군 소속환자들이 약을 꾸준히 복용하지 않는다는 점이 지적되었다.

약을 꾸준히 복용하는 습관이 중요하다?

- 약을 꾸준히 복용하는 사람(adherer)과 그렇지 않은 사람을 나누어서 임상시험결과를 보고한 결과는 다음과 같았다.

*여기서 약을 꾸준히 복용한 사람은 80%이상 약을 복용한 사람을 말한다

	실험군		대조군	
	환자수	사망률	환자수	사망률
Adherers	708	15%	1,813	15%
Non-adherers	357	25%	882	28%
합계	1,103	20%	2,789	21%

[심근경색약(colfibrate) 임상시험 5년 추적 관찰 후 각 그룹별 사망률]
(Statistics, 2007, p. 14)

기타 출처 #1

약을 꾸준히 복용하는 습관이 중요하다?

- **실험군과 대조군사이에 사망률은 차이가 없지만 양 그룹 모두 약을 꾸준히 복용하는 사람들의 사망률이 낮았다!**
- **이 연구는 임상시험이기 때문에 실험연구이지만 약을 복용하는 습관 여부와 사망률의 관계에 관한 부분은 관측연구이다.**
- **이 연구의 결론은**

  1. Colfibrate는 효능이 없다(실험연구 결과).
  2. 약을 꾸준히 복용하는 사람과 그렇지 않은 사람들은 여러 가지 면에서 다르다(관측연구 결과).

- **약을 꾸준히 복용하는 사람들이 일반적으로 건강관리에 신경을 쓰는 편이기 때문에 보이는 현상일 수 있다.**

버클리 대학은 대학원 입시에서 성차별을 하였나?

- 1970년대 미 버클리대학은 대학원 입시에서 성차별이 있었다는 문제 제기를 받게 된다. 그해 대학원 입시 결과에서 남학생들의 합격률은 44%, 여학생들의 합격률은 30%였다.
- 대학원 입시는 각 학과에서 관리를 하기 때문에 대표적인 6개 전공에서 남학생과 여학생의 합격률을 비교한 결과는 다음과 같았다.

버클리 대학은 대학원 입시에서 성차별을 하였나?

	남성		여자	
학과	총지원자	합격률	총지원자	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	6%	341	7%
합계	2,691	45%	1,835	30%

[버클리대학의 6개 주요 단과대학의 성별 대학원 입학자료]
(Statistics, 2007, p. 18)

기타 출처 #2

심슨의 역설(Simpson's Paradox)

- 전체합격률은 남학생이 높지만, 학과별 합격률은 여학생이 높다!
- 이 이유는 남학생들은 다수가 합격률이 높은 A, B학과에 지원하였지만, 여학생들은 극소수가 이들 학과에 지원하여 여학생 전체 합격률은 주로 C, D, E, F학과 합격률에 의해 결정되었고 남학생들의 전체합격률은 A, B학과의 높은 합격률에 힘입어 상승하였다.
- 이렇게 제3의 요인으로 전체자료를 세분화했을 때 정반대의 결과가 나오는 것을 **심슨의 역설(Simpson's paradox)**라고 한다.
- 이 예제에서 주목할 점은 학과와 합격률, 그리고 학과와 성별 지원자 숫자 사이에 강한 연관성이 있어서 전체 성별 합격률에 영향을 미쳤다는 점이다.

교락 효과(Confounder Effect)

- **교락 요인(confounder)는 앞의 예제에서 같이 반응변수(합격 여부)와 설명변수(성별)에 모두 영향을 미치는 변수(학과)를 말한다.**
- **교락 요인을 통제하여야만 반응변수가 순수하게 설명변수에 미치는 영향을 알 수 있다.**
- **교락 요인을 통제하기 위해서**

  1. 교락 요인의 값에 따라 그룹을 나눈 후 반응변수와 설명변수의 관계를 알아본다(subgroup analysis).

  2. 가중평균을 사용한다.

가중평균을 이용한 교락 효과 통제

- 남성합격률을 계산하는 가중평균식은 다음과 같다.

$$\frac{0.62 \cdot 933 + 0.63 \cdot 585 + 0.37 \cdot 918 + 0.33 \cdot 792 + 0.28 \cdot 584 + 0.06 \cdot 714}{4526} = 0.39$$

학과	남성지원자	남성합격률	총지원자
A	825	62%	933
B	560	63%	585
C	325	37%	918
D	417	33%	792
E	191	28%	584
F	373	6%	714
합계	2,691	45%	4,526

[버클리대학의 6개 주요 단과대학의 남성지원자 대학원 합격률]
(Statistics, 2007, p. 18)

기타 출처 #3

가중평균을 이용한 교락 효과 통제

- 즉 전체 지원자 중 그 학과에 지원하는 사람의 비율을 가중치로 한 후 가중치를 학과별 합격률에 곱하여 남학생의 가중합격률 39%를 구하였다.
- 만약 가중치를 그 학과에 지원하는 남학생의 비율로 정한다면 원래 전체합격률인 45%가 계산되어 나온다.
- 즉 학과별로 지원하는 남녀지원자 숫자가 다르다는 점을 통제하기 위해 성별 지원자 비율이 아닌 전체 비율을 가중치로 사용하는 것이다.
- 여학생의 경우 같은 방법으로 가중치를 구하면 43%가 나온다.

오늘의 강의 요점

- 실험 vs 관측연구
- 전향적 연구 vs 후향적 연구
- 연관성 vs 인과성
- 심슨의 역설
- 교락 효과 통계 방법
  - Subgroup analysis
  - 가중평균

출처

#1~3 Freedman, D., Pisani, R. And Purves, R. (2007). Statistics, 4rh edition. W. W. Norton & Company