데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

부분에서 전체를 추론하기

1. 모집단과 표본

영국인의 실제 성관계 상대 수는 몇 명인가?

- 영국의 성생활에 관한 설문조사에서 성관계 상대 수에 관한 내용은 몇 가지 의문점을 포함하고 있다.
- 이 설문조사의 궁극적인 목적은 영국 국민 전체의 성생활 패턴을 알고자 하는 데 목적이 있다.
- 설문조사 결과를 바탕으로 어떻게 영국 국민 전체에 관한 내용을 파악할 수 있을까?
- 이 질문은 다음과 같은 4단계 과정을 거쳐서 대답할 수 있다.

영국인의 실제 성관계 상대 수는 몇 명인가?

- 데이터(Data):
  설문조사 참가자들이 보고한 성관계 상대 수
- 표본(Sample):
  설문조사 대상이 된 영국인의 **실제** 성관계 상대 수
- 연구모집단(Study Population):
  설문조사에 포함될 가능성이 있는 모든 사람들의 성관계 상대 수
- 목표모집단(Target Population):
  영국인 전체의 성관계 상대 수

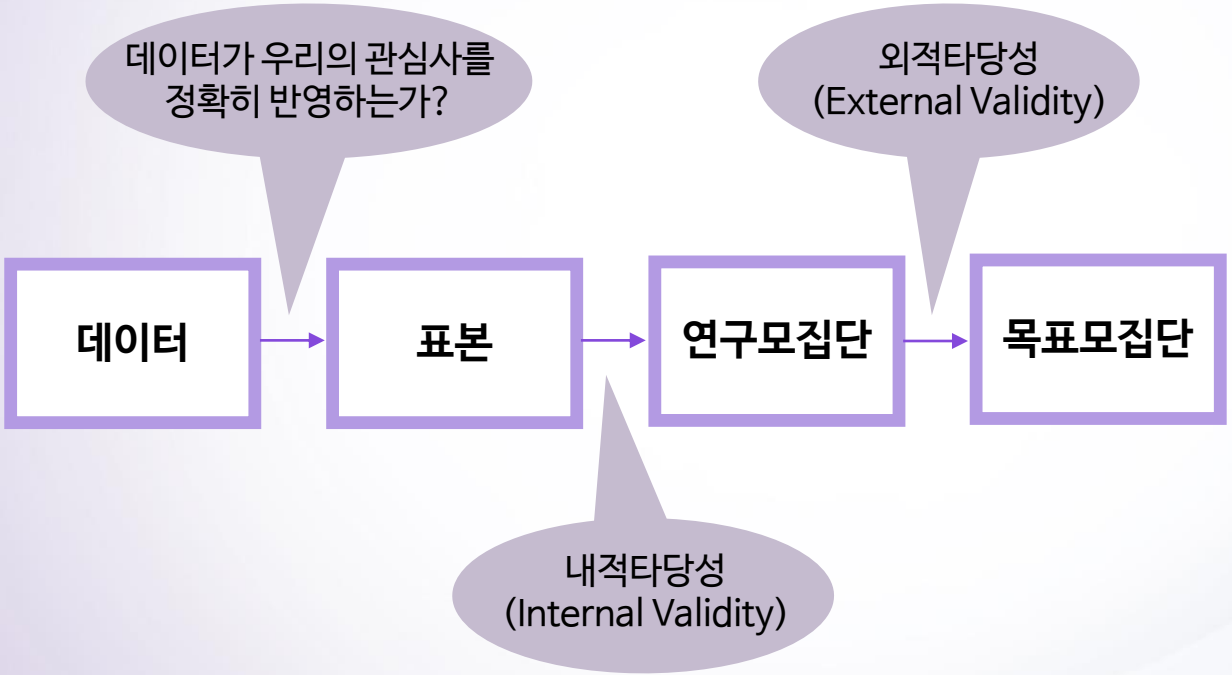
영국인의 실제 성관계 상대 수는 몇 명인가?

- 1단계에서 2단계로 넘어가기 위해서는 설문조사에 사람들이 정직하게 답변한다는 가정이 숨어있다. 하지만 성생활과 같은 민감한 주제에 대해서는 사람들이 거짓말을 할 가능성은 항상 존재한다.
- 2단계에서 3단계로 가는 것은 생각보다 쉽지 않다. 연구모집단의 구성원들이 실제 조사에서 임의로 뽑힐 수 있다는 가정이 성립해야 한다. 영국 성생활 설문조사의 경우 응답률이 약 66% 정도로 높은 편인데 성적 활동이 활발하지 않은 사람은 참가율이 조금 낮다. 하지만 성적으로 자유분방한 사람들 역시 설문조사에 포함하기 어렵다는 점을 고려한다면 큰 문제는 없어 보인다.

영국인의 실제 성관계 상대 수는 몇 명인가?

- 3단계에서 4단계로 가기 위해서 3단계의 연구모집단이 목적모집단의 대표성을 가지고 있어야 한다. 예를 들면 영국 성생활 설문조사에서 교도소, 수도원, 군대와 같은 시설에 있는 사람은 설문조사 대상에 포함되어 있지 않지만, 영국 전체 국민의 성생활 실태를 추론하는 데는 지장이 없다는 가정이 성립하여야 한다.

데이터로 배우기: 귀납적 추론



데이터로 배우기: 귀납적 추론

- **데이터가 표본으로 가는 단계에서 데이터가 가져야 할 특성은 다음과 같다.**

  ⮡ 데이터 자체의 변동이 작고 반복 가능하다.

  ⮡ 알고자 하는 항목에 대해 어떤 편의도 없이 정확히 측정하고 있다.

- **예를 들면 같은 내용을 알고자 하는 설문조사에서 설문을 어떻게 구성하는가에 따라 답이 달라진다면 위의 첫 번째 특성인 반복 가능의 원칙에 어긋나는 것이다. 예를 들면 “참정권 확대를 위해 선거연령을 낮추는데 동의하는가?”와 “학습권을 침해할 우려가 있는 고등학생에게도 선거권을 주어야 하는가?”는 같은 내용에 관한 일반 국민들의 의사를 물어보지만 설문조사 결과는 판이할 수 있다.**

데이터로 배우기: 귀납적 추론

- 표본이 연구모집단의 대표성을 가질 경우 **내적타당성**을 지닌다고 한다. 즉 임의추출과 같은 방법으로 표본을 뽑아서 연구모집단의 대표성을 유지하도록 한다.
- 연구모집단과 목적모집단이 정확히 일치하지 않을 경우 연구모집단의 결과를 목적모집단으로 확장할 수 있는 경우 **외적타당성**을 가지고 있다고 한다. 예를 들면 성인 남성(연구모집단)을 대상으로 신약에 대한 임상시험을 진행한 결과를 전 국민에 대한 결과로 확대 해석하는 경우를 생각해 볼 수 있다.

베트남전 추첨식 징집

[Vietnam Lottery Draft]
(Wikipedia)

- 1969년 미국이 베트남전 참전으로 인해 모병제를 징병제로 바꾸면서 징병 대상 청년들을 대상으로 추첨식으로 징병 순서를 정하였다.
- 투명한 드럼 안에 총 366개의 캡슐이 들어가고 각 캡슐에는 생일이 적혀있었다.

기타 출처 #1

베트남전 추첨식 징집

[Vietnam Lottery Draft]
(Wikipedia)

- 캡슐을 뽑힌 순서로 그 날짜에 태어난 사람들이 징병이 되는 방식이었다. 처음 195개의 캡슐에 해당하는 생일을 가진 사람들이 최종적으로 징집되었다.
- 하지만 추첨 결과 12월 생은 26개의 날짜에 해당되는 사람이 징집된 반면에 1월 생은 14명만 징집되었다.

기타 출처 #1

베트남전 추첨식 징집

- 캡슐은 잘 섞이지 않았고 일반적으로 캡슐을 뽑는 사람들은 위에 놓인 캡슐을 뽑는 경향이 있었다!
- 이런 문제를 해결하기 위해 1970년에 실시한 1951년생 대상 추첨식 징병제에는 2개의 드럼을 준비하고 하나의 드럼에는 생일을, 또 다른 드럼에는 1~365 사이의 숫자를 넣은 365개의 캡슐을 집어넣었다.
- 드럼을 충분히 많이 돌린 후에 먼저 생일이 있는 캡슐을 뽑은 후에 다른 드럼에서 숫자를 뽑으면 그 숫자가 뽑힌 생일의 입대 순서가 되는 방식이었다.

영국에서는 얼마나 많은 범죄가 일어나는가?

- **위의 질문에 답변하기 위해서 두 가지 데이터를 고려할 수 있다.**
  - 영국-웨일스 범죄 설문조사
  - 경찰 범죄보고서

영국에서는 얼마나 많은 범죄가 일어나는가?

- 설문조사의 경우 응답자가 진실을 말하지 않을 경우(1단계 → 2단계), 표본이 실제로 연구모집단을 대표하는지 여부(2단계 → 3단계), 또한 연구모집단에는 16세 미만과 공동시설 거주자들이 제외되어 있다는 사실(3단계 → 4단계)을 고려해야 한다.

- 경찰 범죄보고서의 경우 표본이 연구모집단과 동일하지만 피해자가 보고하지 않은 범죄 혹은 경찰이 기록하지 않은 범죄가 있기 때문에 외적타당성을 가지고 있지 않다. 결론적으로 경찰 범죄보고서는 이러한 문제 때문에 영국에서 국가지정통계 목록에서 제외되었다.

오늘의 강의 요점

- 모집단과 표본
- 귀납적 추론의 4단계: 데이터 → 표본 → 연구모집단 → 목적모집단

출처

#1 https://en.wikipedia.org/wiki/Draft_lottery_(1969)#/media/File:1969_draft_lottery_photo.jpg

부분에서 전체를 추론하기

2. 표본조사방법

센서스(Census)

- **표본을 선택하는 대신 전체 모집단에 대해서 조사를 한 경우를 센서스라고 한다. 대표적인 예로 우리나라 통계청에서 5년마다 실시하는 인구주택총조사를 들 수 있다.**
- **센서스에 대한 문제점으로는**
  - 센서스에 잡히지 않는 사람이 있으며 실제로 이런 사람들은 특정 집단(불법체류자 등)에 속하는 경우가 많다.
  - 모집단은 계속 변하고 있기 때문에 센서스 기간을 고려할 때 완벽하게 모집단의 모든 사람을 조사하는 것은 불가능하다.
  - 센서스가 샘플링보다 복잡할 수 있다.

표본 편의(Sampling Bias)

[『국립공원의 날』 지정 온라인 설문조사]
(국립공원관리공단)

- 무응답 편의(Non-response Bias): 임의로 뽑힌 사람 중 일부만 대답하는 경우 전체모집단을 대표한다고 할 수 없다.

기타 출처 #1

표본 편의(Sampling Bias)

[『국립공원의 날』 지정 온라인 설문조사]
(국립공원관리공단)

- 자원응답편의(Voluntary Response Bias): 원하는 사람만 답변을 한 경우 전체를 대표한다고 하기 힘들다. 예를 들면 웹 설문조사나 학생들이 개설한 자체 강의 평가 웹사이트를 들 수 있다.
- 선택 편의(Selection Bias): 목표모집단과 연구모집단이 상이한 경우 생기는 문제

기타 출처 #1

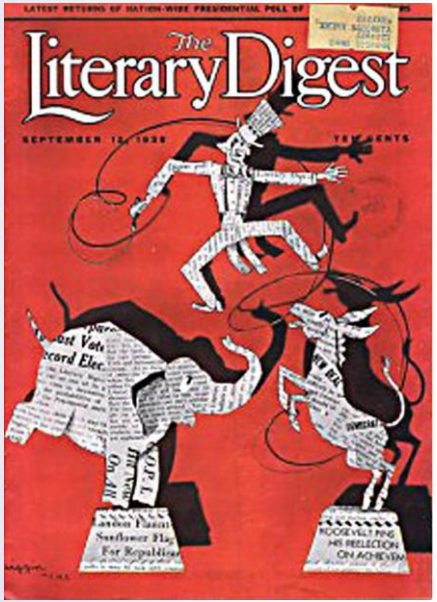
1936년 미국 대통령 선거 예측

[리터러리 다이제스트 표지]
(The Literary Digest, 1936)

- 선택 편의의 대표적인 사례는 리터러리 다이제스트의 1936년 미국 대통령 선거 예측을 들 수 있다.

기타 출처 #2

1936년 미국 대통령 선거 예측



[리터러리 다이제스트 표지]
(The Literary Digest, 1936)

- 리터러리 다이제스트는 천만 명 대상의 설문조사 결과 240만 명에게 답변을 받았고, 그 결과를 바탕으로 공화당 후보인 랜던이 선거를 이길 것으로 예측했으며 민주당 후보인 루즈벨트는 43%만 득표할 것으로 예상하였다.
- 하지만 실제 선거결과는 루즈벨트가 62%의 득표로 압승을 거두게 된다.

기타 출처 #2

1936년 미국 대통령 선거 예측

- **리터러리 다이제스트의 조사대상은 다음과 같았다.**
  - 독자
  - 자동차 소유자
  - 전화번호 소유자
- **대공황 시기 위의 그룹은 고소득층에 속했으며, 고소득층은 공화당을 압도적으로 지지하는 경향이 있었다.**

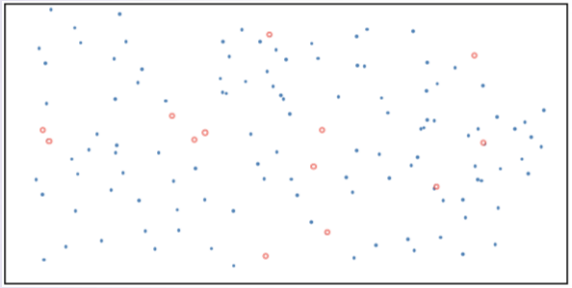
1936년 미국 대통령 선거 예측

- 이 예측 실패를 계기로 리터러리 다이제스트는 폐간의 수순을 밟게 된다.
- 하지만 같은 선거에 대해서 1,500명의 작은 표본으로 정확한 예측을 한 회사가 있었는데 이 회사가 오늘날 여론조사 전문 기관으로 유명한 갤럽이다!

전통적인 표본조사 방법

- 모든 연구에는 모집단의 대표성을 가질 수 있도록 임의표본추출이 가정된다.
- 전통적인 임의표본추출 방법은 다음과 같다.
  - 단순임의추출(Simple Random Sampling)
  - 층화추출(Stratified Sampling)
  - 집락추출(Cluster Sampling)
  - 다단계추출(Multistage Sampling)

단순임의추출(Simple Random Sampling)

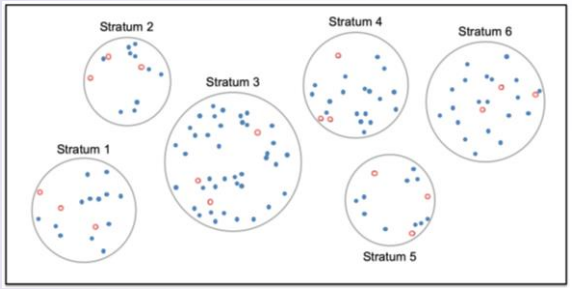


[단순임의추출]
(OpenIntro Statistics, p26)

- 모집단에서 임의로 n개의 표본을 추출할 때 각 표본이 추출될 확률이 모두 동일한 확률이 되도록 추출하는 방법
- 모집단이 큰 경우 비효율적

기 타 출 처 # 3

층화추출(Stratified Sampling)

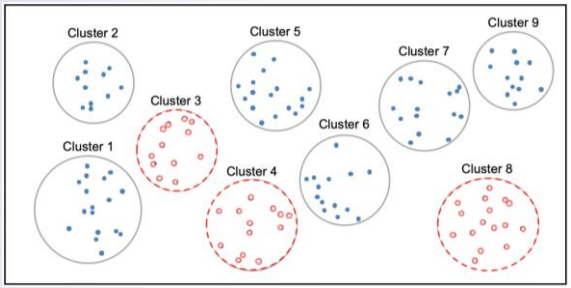


[층화추출]
(OpenIntro Statistics, p26)

- 비슷한 관측치로 이루어진 층(strata)를 만들고 각 층에서 임의로 표본을 추출하는 방법이다.
- 층안은 동질적이고 층 사이는 이질적으로 만들어야 한다.
- 예를 들면 같은 선거구 안에서 특정 동네별로 투표성향이 판이하다면 동네를 층으로 만들어서 표본을 추출한다.

기타 출처 #4

집락추출(Cluster Sampling)

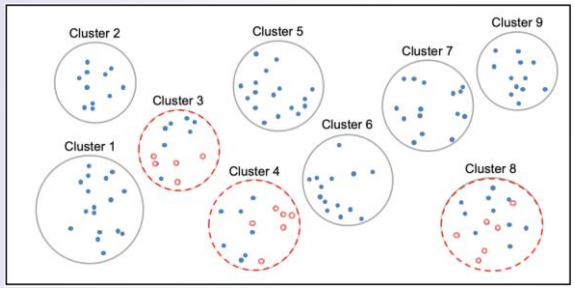


[집락추출]
(OpenIntro Statistics, p28)

- 모집단을 몇 개의 집락(cluster)으로 나눈 후 집락 가운데 몇 개의 집락을 단순임의추출로 추출한 후 추출된 집락 안의 자료를 모두 표본을 간주하는 방법이다.
- 집락 간의 비슷하지만 개개의 집락은 모집단 전체의 특징을 반영할 수 있어야 한다.
- 전쟁, 기근, 자연재해에서 사망률 추정에 사용된다.

기타 출처 #5

다단계추출(Multistage Sampling)



[다단계추출]
(OpenIntro Statistics, p28)

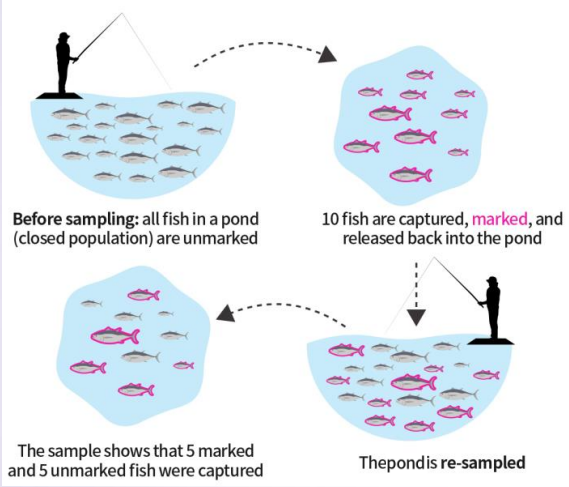
- 집락추출과 동일하게 각 집락을 추출하고 집락안의 자료를 대상으로 단순임의추출을 하여 최종표본을 선정한다.
- 미 갤럽에서 여론조사를 위해 각 지역 번호(cluster)를 임의로 선정하고, 그 지역 번호를 가진 사람을 다시 임의로 선출한다.

기타 출처 # 6

그 외 표본조사방법

- 포획-재포획 추출(Capture-Recapture Sampling)
- 트란섹트 추출(Transect Sampling)

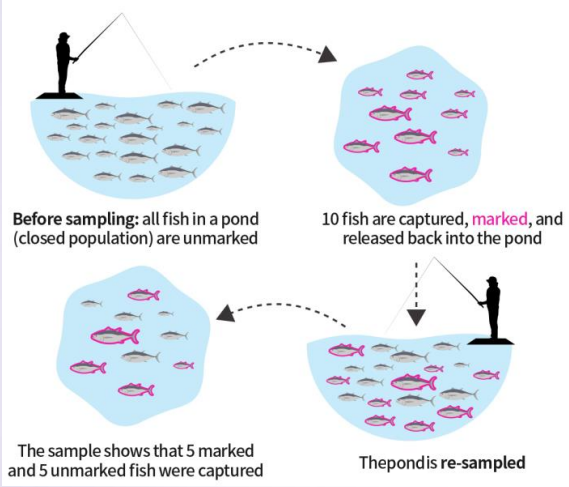
포획-재포획 추출(Capture-Recapture Sampling)



[포획-재포획 추출을 이용한 연못 속 물고기 개수 추정]

- 특정 장소에 살고있는 동물들의 개체 수를 추정하기 위해 사용하는 방법으로 다음과 같은 절차를 사용한다.

포획-재포획 추출(Capture-Recapture Sampling)

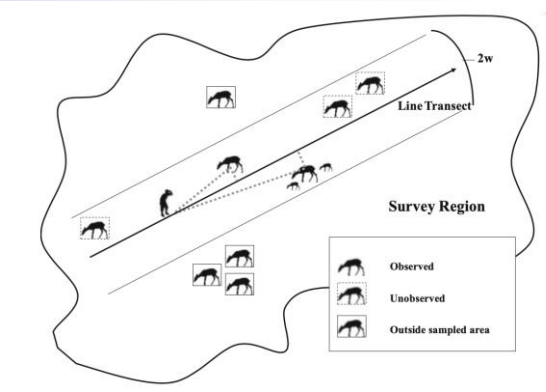


[포획-재포획 추출을 이용한 연못 속 물고기 개수 추정]

- n개의 동물들을 포획한 후 각 동물마다 표지를 부착한다.
- 동물들을 다시 원 서식지로 방목한 후 다시 K개의 동물을 포획하고, 그 중 표지가 있는 동물 개수를 k라고 하자.
- 이 경우 전체 모집단의 개수

$$N = \frac{K \cdot n}{k}$$

트란섹트 추출(Transect Sampling)



[트란섹트 추출]
(Methods for Monitoring Tiger and Prey Populations, p91)

- 캥거루, 고래 등과 같은 야생동물의 숫자를 추정하기 위해 사용하는 방법이다.
- 주어진 공간에 등간격으로 직선을 그은 후, 그 직선을 따라 이동하면서 관측된 동물 숫자를 직선과 동물의 위치와의 수직거리별로 정리를 한 후 그 자료를 이용하여 전체 동물 개수를 추정한다.

기 타 출 처 # 7

오늘의 강의 요점

- **센서스**
- **표본 편의: 무응답 편의, 자원응답편의, 선택 편의**
- **표본조사 방법**
  - 전통적인 표본조사 방법: 단순임의추출, 층화추출, 집락추출, 다단계추출
  - 그 외의 표본조사 방법

출처

#1 국립공원관리공단 http://www.knps.or.kr/portal/events/poll/poll_20180528_1.do

#2 리터러리 다이제스트 표지 https://bit.ly/34iVBjY

#3~6 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

#7 Strindberg, S., Kumar, N. S. Thomas, L. And Goswami, V. R. (2017) Concepts: Estimating Abundance of Prey Species Using Line Transect Sampling.

InL Karanth K., Nichols, J. (eds) Methods for Monitoring Tiger and Prey Populations, Spinger

부분에서 전체를 추론하기

3. 개인정보보호

넷플릭스 경진대회

오늘 밤은 무슨 영화를 볼까?

NETFLIX

- 넷플릭스 가입자들의 영화평에 기초하여 보다 효율적인 추천시스템을 개발하기 위한 데이터 분석 경진대회

- AT&T 연구소의 통계학자들이 주축인 된 BellKor's Pragmatic Chaos 팀 이 우승하여 백만 불의 상금을 받음

	명량	국제시장	부산행	아바타
A	3	4	1	4
B		5		
C	5	2		
D	2		4	?

기타 출처 #1

넷플릭스 2차 경진대회 취소

개인정보보호 누출!

- 1차 경진대회의 성공에 고무된 넷플릭스에서 보다 향상된 추천시스템을 만들기 위해 추가 개인정보를 포함한 2차 경진대회를 추진
- 텍사스 대학의 연구진에 의해 IMDb라는 또 다른 영화 별점 사이트의 DB와 연동할 경우 개인신상을 알아낼 수 있다는 점이 지적됨
- 이러한 개인정보 누설에 대한 우려로 2차 경진대회 취소

기타 출처 #1

데이터 3법이란?

국내 개인정보 관련 보호 법률 현황

- 데이터 3법(개인정보보호법, 정보통신망법, 신용정보법) 개정안을 통해 개인정보보호법에서 가명정보 소개
- 가명정보의 경우 개인동의 없이 통계작성, 연구, 공익적 기록보존 목적으로 사용가능

개인정보/가명정보/익명정보

	개념	예	활용가능범위
개인정보	특정개인에관한정보	한석규, 1964년11월3일생 2020년5월1일, 왓챠에서〈명량〉시청	사전적이고구체적인 동의후에활용가능
가명정보	추가정보없이 특정개인을알아볼수없게 처리한정보	한XX, 1964년생 2020년5월1일, 왓챠에서〈명량〉시청	통계작성/연구/ 공익적기록보존의경우 동의없이활용가능
익명정보	복원이불가능할정도로개 인을알아볼수없게처리한 정보	남성,50대 2020년5월1일, 왓챠에서시대물시청	제한없이활용가능

데이터 거래소

- **데이터 3법의 통과와 함께 다양한 형태의 데이터 거래소 등장**
  - 금융데이터거래소
  - 교통데이터거래소
  - 민간데이터거래소

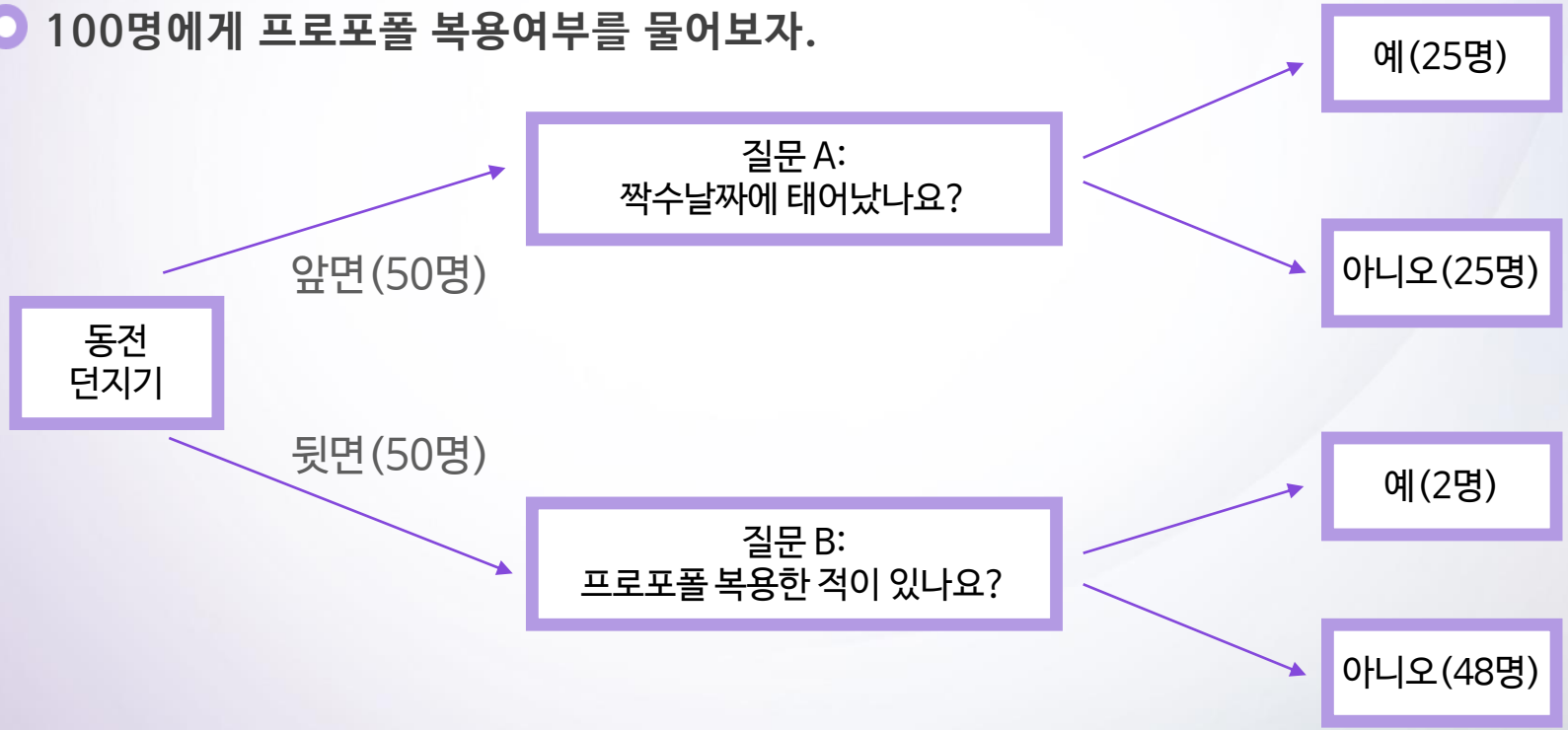
뉴욕 시 택시 자료

- 2015년 뉴욕 시 택시 및 리무진 위원회에서 2009년 1월부터 2015년 6월까지 약 11억 건의 뉴욕 시 택시 운행 자료공개
- 파파라치의 사진을 통해 미국 유명 여배우 제시카 알바가 택시 기사에게 팁을 주지 않았음이 알려짐

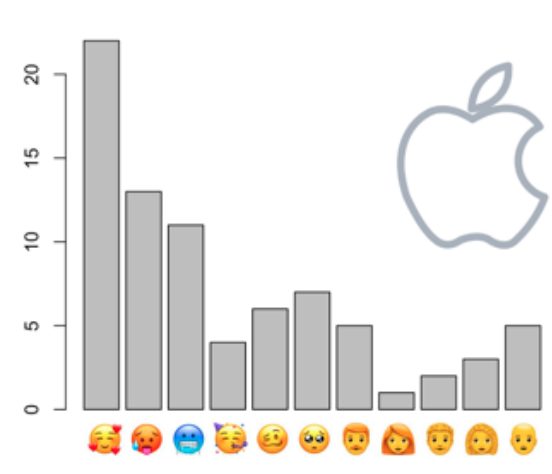
국소차등정보

프로포폴 복용한 적이 있나요?

- 100명에게 프로포폴 복용여부를 물어보자.



애플에서 국소차등정보보호 활용



[이모티콘 사용빈도]

- 애플에서 서비스 개선을 목표로 이모티콘 사용현황을 파악하고자 함
- 개별 이모티콘 사용내역은 국소차 등정보보호를 통해 실제 애플은 각 이모티콘의 전체 사용빈도는 파악하지만 개인별 이모티콘 사용현황은 알지 못함
- 이모티콘보다 더욱 민감한 위치정보 등에 국소차등정보보호 적용 가능

기타 출처 #2

통계청 마이크로 데이터 센터

- 정부각부처와 지자체, 연구기관이 생성하는 마이크로 데이터를 한 곳에 모아 이용할 수 있도록 서비스하는 센터
- 2019년 현재 총 10개의 Research Data Center를 운영하여 이용 센터 서비스 제공

  ↳ 이용 센터 서비스: 지정된 장소에서 제공받은 자료를 분석하고 결과만 승인 하에 반출하는 서비스

건강보험심사평가원 코로나-19 데이터 공유센터

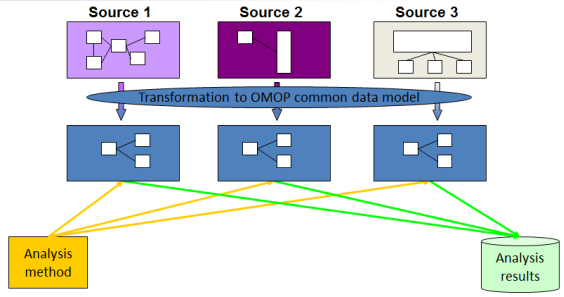
[HIRA COVID-19]
(건강보험심사평가원)

[HIRA COVID-19 (CDM)]
(OMOP)

- 3월 27일 보건복지부와 건강보험심사평가원에서 익명화된 국내 코로나-19 환자 데이터 공개
- 건강보험심사평가원이 보유한 전국민 청구데이터를 근간으로 웹사이트에 샘플데이터를 공통데이터 모델 형식으로 공개

기타 출처 #3, 4

공통데이터모델이란?



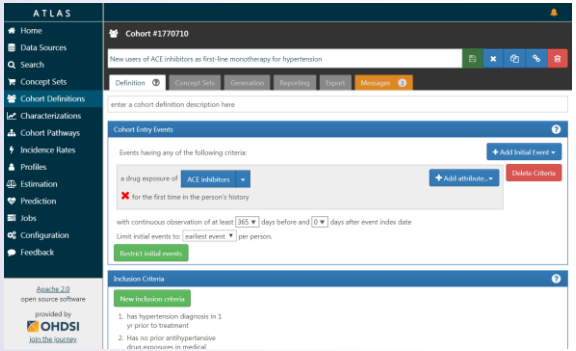
[OMOP Common Data Model]
(Observational health Data Sciences and Information)

- 서로 다른 데이터 생성기관이 데이터의 직접 공유대신 데이터 표준화를 통하여 분석결과를 공유하는 방식
- 의료기관을 중심으로 사용되고 있으며 OHDSI 국제 컨소시움의 경우 전세계 14개국 200개 이상의 기관이 참여하고 있음

기타 출처 #5

공통데이터모델의 분석

ATLAS란?



[ATLAS]
(Observational health Data Sciences and Information)

- OHDSI에서 개발된 웹 기반 무료분석 툴
- 분석 매뉴얼과 튜토리얼 영상을 무료로 제공
- 이 외에도 시각화 툴 Achilles와 R 패키지 모음 HADES (Health Analytics Data-to Evidence Suite) 제공

기타 출처 # 6

오늘의 강의 요점

- 우리, 개인정보 알아냈을까? – 차등정보보호가 있잖아
- 데이터 공유 못하지만 괜찮아! – 공통데이터모델을 사용하면 되잖아

출처

#1 Netflix https://www.netflix.com/kr/

#2 Copyright 2020. 장원철 all right reserved

#3 건강보험심사평가원 http://www.hira.or.kr/main.do

#4~6 Observational health Data Sciences and Information https://www.ohdsi.org/

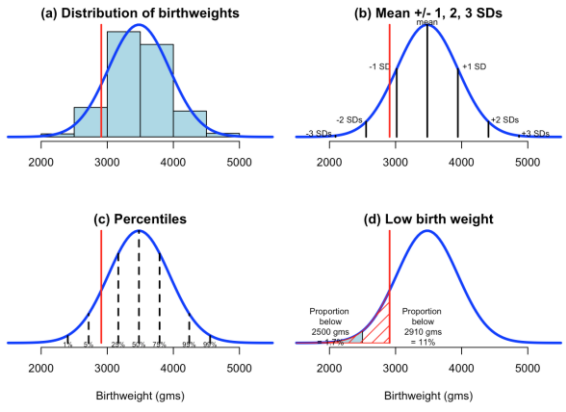
부분에서 전체를 추론하기

Lab 4: 정규분포

모집단 분포(Population Distribution)

- 미국에 있는 한 (히스패닉이 아닌) 백인 여성이 2.91kg의 아기를 낳았다. 그렇다면 이 몸무게는 비정상적으로 낮은가?
- 위의 질문에 답변하기 위해서 최근 출산한 신생아 전체의 몸무게 분포에서 2.91kg이 위치를 알아보고자 한다. 미국에서는 인종별로 출생 체중을 미국 인구동태 통계 시스템에 보고하게 되어 있는데 백인 신생아의 평균 체중은 3,480g으로 알려져 있다.
- 여기서 백인 신생아를 모집단으로 볼 수 있고 몸무게, 키 등과 같이 여러 가지 요인들에 의해서 결정되는 측정치의 분포는 종모양 곡선(bell shaped curve)을 따른다고 알려져 있다.

정규분포(Normal Distribution)

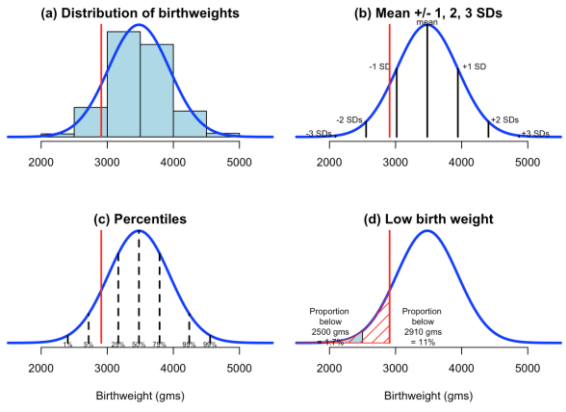


[미국 만삭(39-40주)인 백인 산모가 낳은 신생아 1,096,277명의 분포]
(The Art of Statistics, p87)

- 종 모양 곡선은 정규분포(Normal Distribution)라는 이름으로 잘 알려져 있다.
- 정규분포는 1809년 요한 카를 프리드리히 가우스가 천문학과 인구조사의 측정오차를 다루는 과정에 유래한다.
- 그림(a)는 2013년에 백인 산모에게서 태어난 신생아 1,096,227명의 자료를 히스토그램과 정규분포를 이용해서 나타낸 그림이다.

기타 출처 #1

정규분포(Normal Distribution)

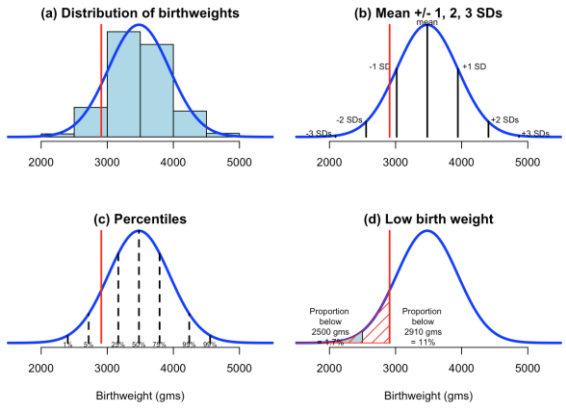


[미국 만삭(39-40주)인 백인 산모가 낳은 신생아 1,096,277명의 분포]
(The Art of Statistics, p87)

- 신생아 체중 자료의 경우 모집단의 평균은 3,480g, 표준편차는 462g이다.
- 이와 같은 모집단의 요약치를 모수(parameter)라고 부르고 정규분포의 경우 2가지 모수를 통해서 분포의 모양이 결정된다.

기타 출처 #1

정규분포(Normal Distribution)



[미국 만삭(39-40주)인 백인 산모가 낳은 신생아 1,096,277명의 분포]
(The Art of Statistics, p87)

- 그림(d)를 통해서 정규분포를 사용한다면 2,910g의 몸무게를 가진 신생아의 경우 11 백분위수(percentile)에 해당한다는 것을 알 수 있다. 즉 2,910g보다 몸무게가 적게 나가는 신생아는 전체 신생아 중 11%이다.

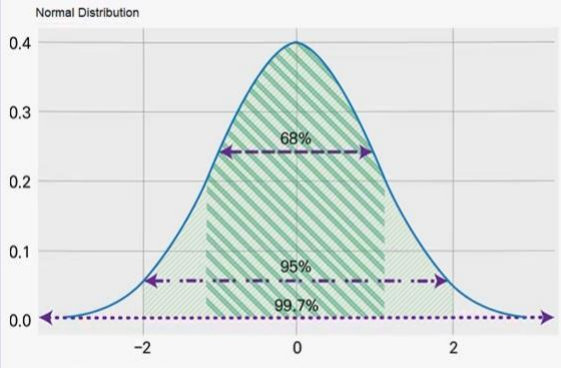
기타 출처 #1

Web Applet을 이용한 백분위수 계산

- Applet:
  http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

정규분포(Normal Distribution)

68-95-99.7 법칙



[68-95-99.7 법칙]

정규분포의 경우

- 약 68%의 자료가 평균으로부터 1 표준편차 범위 안에 있다.
- 약 95%의 자료가 평균으로부터 2 표준편차 범위 안에 있다.
- 약 99.7%의 자료가 평균으로부터 3 표준편차 범위 안에 있다.

기타 출처 #3

표준화 점수

- **학생 A와 B가 수능에서 각각 한국지리와 세계사를 선택하였을 경우 성적을 공정하게 비교할 수 있을까? 두 과목 점수의 모집단분포가 정규분포를 따른다고 가정할 수 있지만 평균과 분산은 다르다.**
- **이 경우 원 점수를 다음과 같은 절차를 거쳐서 평균이 50과 분산이 10인 표준점수로 환산하여 비교한다.**

1. 원 점수를 표준화(과목별 평균을 뺀 후 표준편차로 나누어 주는 작업)을 한다. 표준화된 점수는 평균이 0이고 표준편차가 1인 표준정규분포를 따른다.

2. 표준화된 점수를 평균이 50점이며 표준편차가 10인 표준점수로 변환한다. 이렇게 하기 위해 표준화된 점수에 10을 곱하고 50을 더해준다.

표준화 점수

- 정국은 한국지리에서 46점을, 예린은 세계사에서 48점을 받았다고 가정하자. 한국지리를 택한 전체 수험생 점수의 평균은 40점, 표준편차는 4점이고 세계사의 경우 평균이 42점, 표준편차가 3점이라고 하자. 두명의 표준점수는 구해보자.
- 정국의 표준화 점수는 (46-40)/4=1.5이며 표준점수의 경우 1.5x10+50 =65점이다.
- 예린의 표준화 점수는 (48-42)/3=2이며 표준점수의 경우 2x10+50=70점이다.
- Web Applet을 이용해서 백분위수를 계산하면 정국의 한국지리 점수는 93백분위수, 예린의 세계사 점수는 97%백분위수에 해당한다.

표준화 점수

- **수능에서 과목별 1등급을 받기 위해서는 상위 4% 즉 96백분위수 보다 점수가 높아야 한다. 앞의 예제에서 정국의 점수는 2등급에 해당한다. 세계사에서 1등급을 받기 위해서 점수는 얼마여야 하나?**

- **다음과 같은 2단계 과정을 거쳐서 질문에 답변할 수 있다.**

  - 표준정규분포에서 96백분위수를 먼저 찾는다. 표를 이용하거나(추천하지 않음) 컴퓨터 프로그램(Web Applet)을 이용할 경우 쉽게 1.75임을 알 수있다.

  - 표준점수 기준으로는 1.75x10+50=67.5이며 원점수 기준으로는 1.75x4+40=47임을 알 수 있다.

오늘의 강의 요점

- 모집단 분포
- 정규분포
- 표준화 점수

출처

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

#2 Applet: http://www.socr.ucla.edu/htmls/SOCR_Distributions.html