

1. 데이터과학과 통계

	단위별 학습내용 (Week1)
wk1-1	데이터과학이란 무엇인가
wk1-2	통계가 상식이 된 사회
wk1-3	데이터분석과 윤리
wk1-4	공유데이터와 오픈소스

Wk1-1 : 데이터과학과 통계 - 데이터과학이란 무엇인가 -

통계학

인공지능

데이터마이닝

딥러닝

빅데이터

데이터 (Data)



데이터분석
(Data Analytics)



인사이트 창출
(Insight)



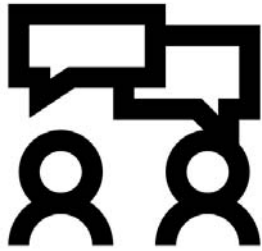
데이터 큐레이션
(data curation)
데이터추출, 변환
- SQL, R, Python

통계모형, 인공지능
(t-검정, 회귀분석, 머신러닝)

데이터 시각화
(data visualization)
그래픽 (R의 ggplot)

1. 데이터과학이란

1.1 데이터과학이란 무엇인가



Analytics



Discovery



Insight

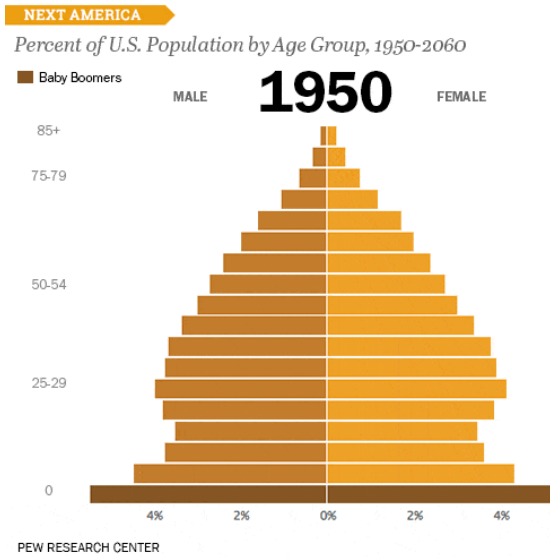
1. 데이터과학이란

1.1 데이터과학이란 무엇인가

- 통계적 개념과 지식 – 샘플링, 확률분포, 가설검정, p-value
- 데이터를 다룰수 있는 기술 (데이터 큐레이션) – 빅데이터 다루기
- 데이터의 요약된 정보 전달 기술 – 데이터 시각화 (공간지도분석, 다차원그래픽)
- 데이터윤리, 데이터보안
- 데이터 도메인에 대한 지식과 분석능력 (현실 문제의 해결능력)

2. 데이터과학의 예시

1.1 데이터과학이란 무엇인가

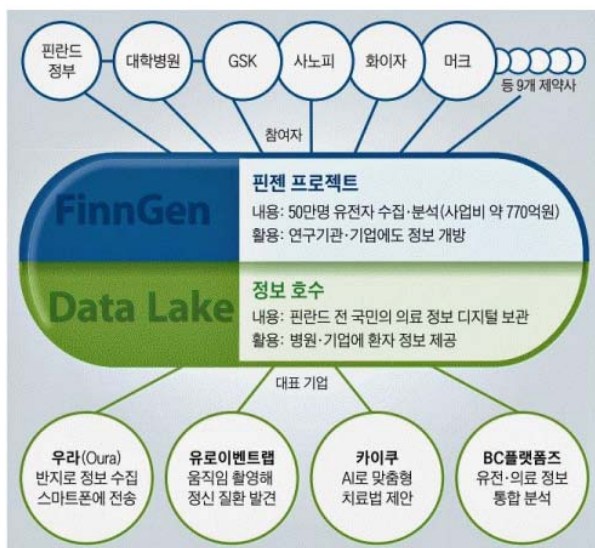


- 데이터 분석 결과를
쉽게 이해할 수 있도록 보여주는 것!
- 그래프, 도표, 이미지, 단어 구름 등을
통해 한 눈에 이해할 수 있도록 하는 것!

2. 데이터과학의 예시

1.1 데이터과학이란 무엇인가

• 핀란드의 의료데이터 프로젝트 (FinnGen)



- 핀란드인(Finnish)과 유전자(Genome)의 합성어.
- 자발적 참여자의 유전자정보를 수집하고 환자의 의료정보까지 통합구축.
- 현재 50만명 목표에서 23만명 수집. 그 중 15만명의 유전자 정보 보유.
- 6개월마다 데이터 업데이트 – 전세계 연구자와 공유
- 관절염/당뇨병 등 자가면역질환 연구 수행중 – 개인 맞춤형 약 개발 추진중

- 데이터과학을 위한 통계적 개념과 지식
- 공유데이터와 오픈소스
- 빅데이터분석을 위한 첫걸음 – 데이터의 중심위치, 산포정도
- 데이터의 시각화
- 데이터과학에서 확률분포는 무슨 의미를 전달



Wk1-2 : 데이터과학과 통계 - 통계가 상식이 된 사회-

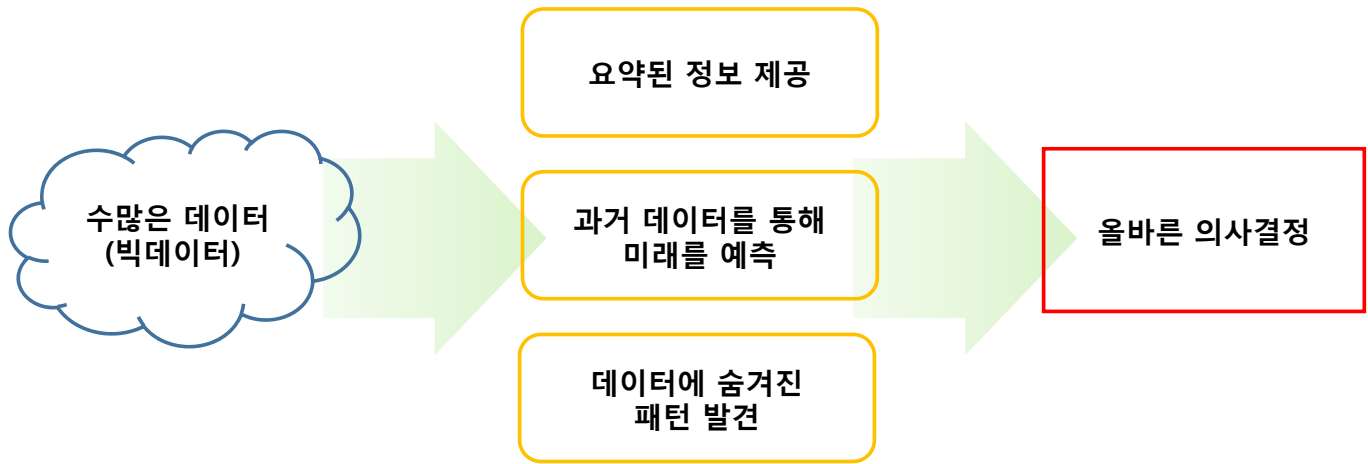
1. 통계가 왜 필요한가

통계가 왜 필요한가?
=
데이터를 올바르게
다룰줄 알면 무엇이 더 좋아지나?

1. 통계가 왜 필요한가

1.2 통계가 상식이 된 사회

- 통계는 올바른 의사결정을 돕는다.



1. 통계가 왜 필요한가

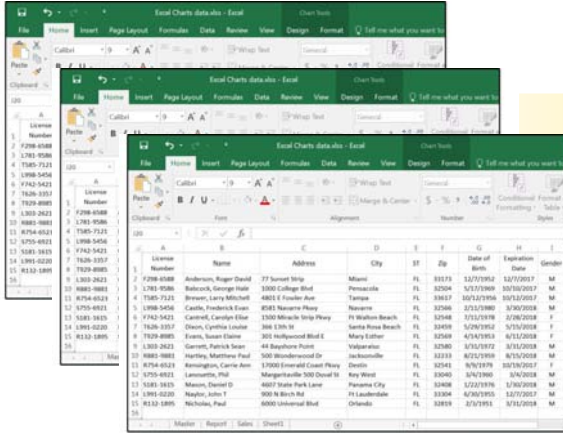
1.2 통계가 상식이 된 사회

요약된 정보 제공

1. 통계가 왜 필요한가

1.2 통계가 상식이 된 사회

세 줄 요약 좀!



License Number	Name	Address	City	ST	Zip	Date of Birth	Expiration Date	Gender
7298-0588	Anderson, Roger David	77 Sunset Strip	Miami	FL	33175	12/7/1957	12/7/2017	M
1785-0586	Balchuck, George Hall	3805 College Blvd	Pensacola	FL	32504	5/23/1969	10/10/2017	M
7185-7321	Brown, Larry Mitchell	4821 E Fowler Ave	Tampa	FL	33617	10/13/1958	10/13/2017	M
1388-5454	Castle, Frederick Evan	8581 Newgate Place	Newport	FL	32566	2/11/1980	3/10/2018	M
7182-1421	Carmel, Candice Ellen	7302 Miracle Strip Pkwy	39 Wablen Beach	FL	32548	7/11/1919	2/10/2018	F
7626-1357	Chen, Cynthia Louise	386 E 9th St	Santa Rosa Beach	FL	32459	5/29/1962	3/7/2018	F
7029-0885	Evans, Susan Elaine	301 Hollywood Blvd E	Mary Esther	FL	32568	4/24/1953	6/12/2018	F
1303-2623	Garnett, Patrick Sean	44 Bayshore Point	Valparaiso	FL	32780	3/10/1972	3/10/2018	M
8881-0882	Hartley, Matthew Paul	500 Woodrowood Dr	Jacksonville	FL	32210	8/21/1959	8/21/2018	M
8774-6121	Hennington, Carrie Ann	17000 Emerald Coast Pkwy	Greenville	FL	32742	8/6/1978	10/16/2017	F
1775-4923	Lewis, Phil	Margueriteville 500 Donald St	Key West	FL	33040	3/4/1960	3/4/2018	M
1581-1015	Mason, Quinn D	4807 State Park Lane	Panama City	FL	32408	1/12/1919	1/12/2018	M
1391-0230	Naylor, John T	500 N Birch Rd	74 Lauderdale	FL	33304	6/26/1953	12/7/2017	M
8112-1895	Nicholas, Paul	6900 Universal Blvd	Orlando	FL	32819	2/17/1951	3/7/2018	M

월 평균 방문객 3000명,
금요일 평균 방문객이 가장 많고
방문객 수는 증가하는 추세

엄청난 데이터가 실시간으로 들어오고 있다고 하면...

1. 통계가 왜 필요한가

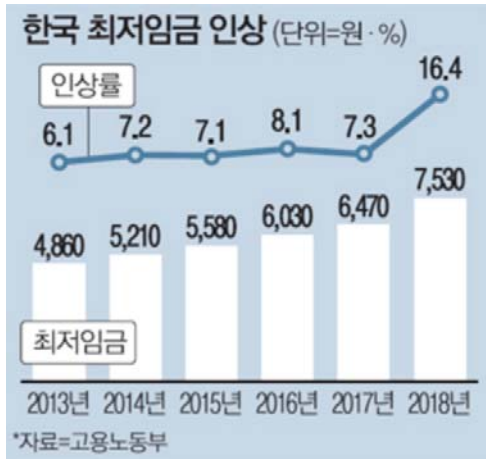
1.2 통계가 상식이 된 사회

과거 데이터를 통해
미래 데이터를 예측

1. 통계가 왜 필요한가

1.2 통계가 상식이 된 사회

과거시점의 데이터로 미래시점을 예측!



지난 6년 간의 변화를 보니
2019년 최저임금은
8000~8500원 정도 !

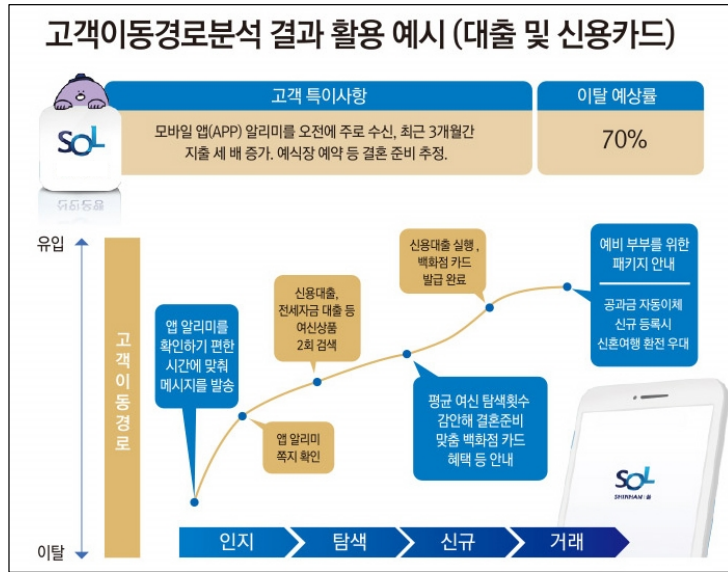
2. 의사결정에서 통계치의 역할

1.2 통계가 상식이 된 사회

데이터에 숨겨진
패턴 발견

2. 의사결정에서 통계의 역할

1.2 통계가 상식이 된 사회



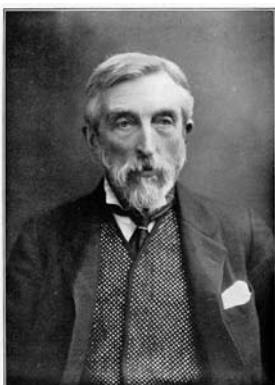
S 은행의 고객 맞춤형 마케팅

- 고객의 금융 검색 경로를 추적
- 고객 이동경로(Customer Journey) 분석
- 이탈 가능성이 높은 고객을 붙잡고 신규 고객 유입하는데 활용

2. 의사결정에서 통계의 역할

1.2 통계가 상식이 된 사회

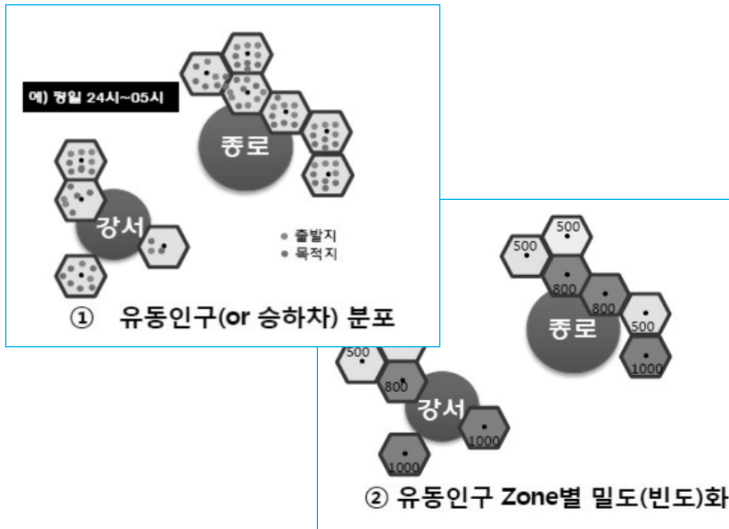
정부 정책의 근거자료 - 영국의회 노령연금 도입



Charles Booth (1840~1916)

- 1886년 영국의 사회학자 찰스 부스(Charles Booth)
- 산업혁명으로 부유해진 런던에서 시민 빈곤 상황을 12년간 조사
- 가난을 8단계로 분류하여 절대빈곤이 30.7%에 달한다는 결과를 발표
- 1908년 영국의회의 노령연금 도입

정부 정책의 근거자료 - 서울시 심야버스 노선정책



서울시 심야버스 심층 분석도

- 시민들이 사용한 자정~새벽 5시까지의 휴대폰 전화 데이터 수집
- 유동인구 분포 및 밀도를 파악하여 유동인구 및 교통수요가 많은 곳에 심야버스 노선 수립

5. 요약

- 개인의 일상활동은 데이터화를 통해 예측분석이 가능하도록 수량화, 객관화되어진다.
- 통계는 수많은 데이터로부터 요약된 정보를 제공, 미래 데이터를 예측, 숨겨진 패턴을 발견함으로써 올바른 의사결정을 돕는다.
- 통계치는 금융권의 관리전략, 정부 정책 수립, 법정소송에서의 근거자료 등으로 활용될 수 있다.



Wk1-3 : 데이터과학과 통계 - 데이터분석과 윤리 -

1. 데이터의 정직성

- 한강 수질 검사를 위해 한강물 채취



한강에서



이만큼만!

- 채취한 한강물을 집에 있던 보온병에 담아 방안에 [보관](#)



보온병에 담아



방안에 보관!

세계 최고 수질 검사 기관에서
분석한 결과,
마셔도 되는 물로 판명됨!

1. 데이터의 정직성

1.3 데이터분석과 윤리

- 왜 신뢰할 수 없을까?
- 데이터를 잘못 수집했다!



한강이 얼마나 넓은데
어디서 수집?

어제 산성비가 내렸을지도
모르잖아!

손으로 채집하면 어떡해!
오염됐을지도 몰라!

1. 데이터의 정직성

1.3 데이터분석과 윤리

- 왜 신뢰할 수 없을까?
- 데이터를 잘못 수집했다!
- 데이터를 잘못 보관했다!



보온병에 있던 세균이 옮으면
어떡해?

차가운데
보관해야하는 거 아냐?

다른 사람이 손대지 못하게
안전한 곳에 보관해야지!

너무 적은 양의 데이터,
편향된 표본 추출,
데이터의 왜곡 및 훼손, ...

- **바르지(정직하지) 못한 데이터**
- **데이터분석은 무의미!**



- 정직하지 못한 데이터의 주요 원인은
데이터 분석가의
 - 비윤리성
 - 무지함
 - 환경의 제약

2. 기사에 등장하는 통계치 해석과 평가

1.3 데이터분석과 윤리

비정규직 임금, 정규직의 54.4% '146만원' ...매년 더 벌어져

[뉴스] 입력 2015.11.04 12:12

4일 통계청이 발표한 '경제활동인구조사 근로형태별 및 비임금근로 부가조사 결과'에 따르면 올 6~8월 정규직 근로자의 월평균 임금은 269만6000원으로 비정규직의 월평균 임금(146만7000원)보다 122만9000원 더 많은 것으로 나타났다. 비정규직의 임금은 정규직의 54.4% 수준이다.

- 동등한 조건으로 비교 필요 (주5일 근무)
- 근로 시간이 적은 비정규직의 월급여가 전일제인 정규직의 월급여보다 적은 것은 당연한 결과
- 성, 연령, 근속년수 등의 요인을 통제한 후 시급으로 비교하는 것이 적합

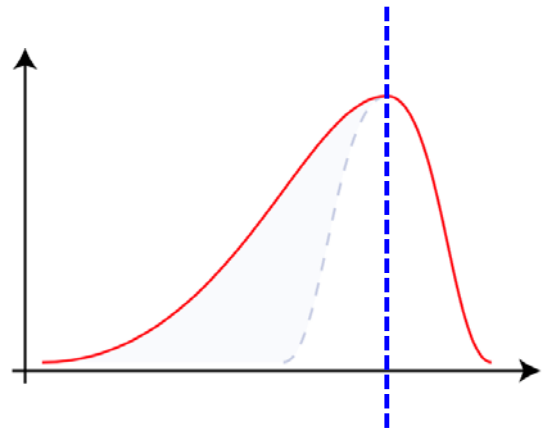
2. 기사에 등장하는 통계치의 해석과 평가

1.3 데이터분석과 윤리

세무사 月 1073만원 최고소득

입력: | 수정: 2009-12-16 01:10

- 편향된 표본 추출
- 알고보니 세무사 39명을 대상으로 조사
- 그중 연봉이 3억~4억원대인 자영업 세무사가 우연히 많았던 것



- **데이터과학의 윤리**는 데이터를 올바르게 분석할 뿐만 아니라 올바른 방법으로 데이터를 수집해야 함을 의미한다.
- **정직하지 못한 데이터의 주요 원인은** 데이터분석자의 무지함, 비윤리성, 그리고 환경의 제약에 의한다.
- **데이터 수집 시** 너무 적은 양의 데이터, 편향된 표본 추출, 데이터의 왜곡 및 훼손에 주의해야 한다.
- **결측치 문제**도 고려해야 한다.



Wk1-4 : 데이터과학과 통계

- 공유데이터와 오픈소스 -

1. 공유데이터란

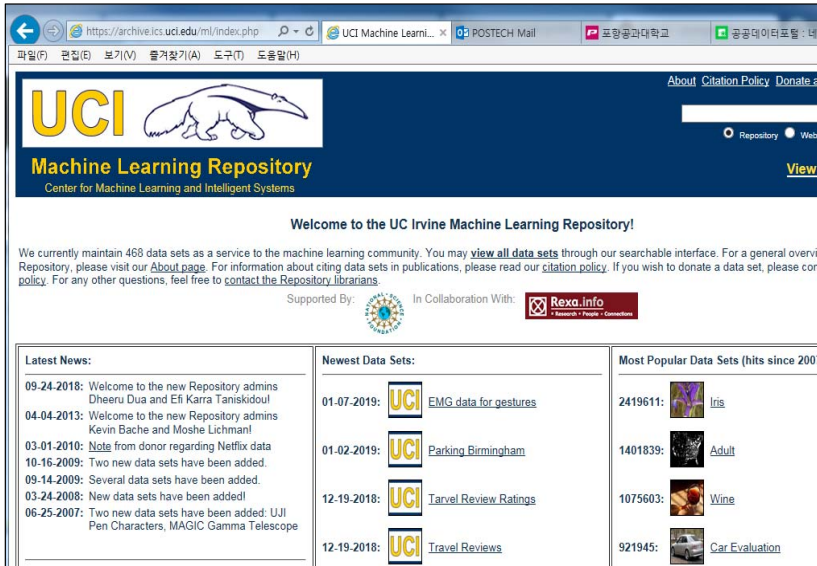


- 모든 사람이 자유롭게 사용 및 재사용
이 가능하며 재배포할 수 있는 데이터
- 이용성 및 접근성
- 재사용과 재배포
- 보편적 참여

1. 공유데이터

1.4 공유데이터와 오픈소스 (github)

• 연구자들을 위한 공유데이터 서비스



머신러닝기법분석에 활용가능한 데이터 저장소
Machine Learning Repository in UC, Irvine

<https://archive.ics.uci.edu/ml/index.php>

1. 공유데이터의 예시

1.4 공유데이터와 오픈소스 (github)



인기 데이터															
<table> <tr> <td>SHEET 4,845 건</td><td>통계 데이터를 포함한 문서, 분석을 정보 등을 SHEET 형태로 제공합니다.</td></tr> <tr> <td>OPEN API 4,718 건</td><td>민간과 개인의 기술 업 개 필요한 정보를 제공</td></tr> <tr> <td>CHART 1,282 건</td><td>통계성 데이터를 CHA</td></tr> <tr> <td>MAP 343 건</td><td>공간정보를 MAP 형태</td></tr> <tr> <td>FILE 904 건</td><td>유동인구, 사업체 정보 등 보고서를 FILE 형태로 제 공합니다.</td></tr> <tr> <td>LOD 208 건</td><td>서울시에서 개방한 공 공데이터를 외부 데이 터와 의미적으로 연 결하는 서비스입니다.</td></tr> <tr> <td>LINK 62 건</td><td>각종 정보 조회, 예약 서비스 등을 LINK 형 태로 제공합니다.</td></tr> </table>	SHEET 4,845 건	통계 데이터를 포함한 문서, 분석을 정보 등을 SHEET 형태로 제공합니다.	OPEN API 4,718 건	민간과 개인의 기술 업 개 필요한 정보를 제공	CHART 1,282 건	통계성 데이터를 CHA	MAP 343 건	공간정보를 MAP 형태	FILE 904 건	유동인구, 사업체 정보 등 보고서를 FILE 형태로 제 공합니다.	LOD 208 건	서울시에서 개방한 공 공데이터를 외부 데이 터와 의미적으로 연 결하는 서비스입니다.	LINK 62 건	각종 정보 조회, 예약 서비스 등을 LINK 형 태로 제공합니다.	<div> <div>파일데이터</div> <div><> 오픈 API</div> </div> <p>무로개방주차장(명절연휴) 현황 수원시 이미용업현황 상가(상권)정보 대전광역시_상수도수질검사 정보 대한무역투자진흥공사 해외진출 한국기업 디렉토리 DB</p>
SHEET 4,845 건	통계 데이터를 포함한 문서, 분석을 정보 등을 SHEET 형태로 제공합니다.														
OPEN API 4,718 건	민간과 개인의 기술 업 개 필요한 정보를 제공														
CHART 1,282 건	통계성 데이터를 CHA														
MAP 343 건	공간정보를 MAP 형태														
FILE 904 건	유동인구, 사업체 정보 등 보고서를 FILE 형태로 제 공합니다.														
LOD 208 건	서울시에서 개방한 공 공데이터를 외부 데이 터와 의미적으로 연 결하는 서비스입니다.														
LINK 62 건	각종 정보 조회, 예약 서비스 등을 LINK 형 태로 제공합니다.														

• 정부에서 제공하는 공공데이터

- 통계청 (kostat.go.kr)
- 공공데이터포털 (www.data.go.kr)
- 서울열린데이터광장 (data.seoul.go.kr)

1. 공유데이터의 예시

1.4 공유데이터와 오픈소스 (github)

네이버의 데이터랩

국내 공공데이터를 기관별로 분류하여 접근성을 높임

The image shows the NAVER DataLab interface. On the left, a sidebar lists various data categories: e-나라지표, 경기도 공공데이터, 공공데이터포털 (highlighted), 국가기록원, 문화재청, 서울열린데이터광장, and 열린재정. The main area displays a list of public data categories with their counts: 전체 31,019, 공공행정 4,531, 과학기술 1,056, 교육 1,600, 교통물류 3,220, 국토관리 1,391, 농축수산 1,324, 문화관광 4,942, 법률 109, 보건의료 2,000, 사회복지 2,116, and 산업고용 2,344. A blue arrow points from the '공공데이터포털' category in the sidebar to a detailed view of the '공공데이터포털 > 공공행정' section. This section lists specific datasets, such as '경상북도 경산시_제조업등록현황' and '부산광역시_북구_사회적기업 현황', each with a date and a brief description.

(<https://datalab.naver.com/opendata.naver>)

2. 오픈소스

1.4 공유데이터와 오픈소스 (github)



- 저작권자가 소스 코드를 공개하여 누구나 복제, 개작, 배포할 수 있는 소프트웨어
- R, Python : 오픈소스 통계분석 프로그램
- C++, 자바, 파이썬 등 다른 프로그래밍 언어와 쉽게 연동
- 빅데이터 시스템인 스파크와도 일부 기능을 연동함으로써 응용범위가 더욱 넓어짐

2. 인공지능에서의 오픈소스

1.4 공유데이터와 오픈소스 (github)



- 구글은 머신러닝과 신경망 연구를 위한 소프트웨어 **텐서플로우**를 오픈소스로 공개
- 구글 딥마인드는 인공지능 개발 플랫폼인 **딥마인드랩**을 공개해 누구나 인공지능 알고리즘을 테스트해볼 수 있게 함

2. 공유데이터 vs. 오픈소스

1.4 공유데이터와 오픈소스 (github)



공유데이터

단순히 수치로 표현되는 측정치
또는 결과 값으로 표현



오픈소스

단순 데이터가 아닌 지적 창작물



- **Git** : 프로그램 등의 소스 코드 관리를 위한 분산 관리 툴 (프로그램 소스를 공유하고 협업하여 개발할수 있는 버전관리 시스템)
- **GitHub**은 Git에 프로젝트 관리지원기능을 확장한 웹 호스팅 서비스. Git을 손쉽게 이용 및 오픈소스 개발자들을 확산하는 데 중요한 역할을 한 웹 서비스

- 2008년 미국 Github사에서 서비스를 시작
- 현재 전세계에서 오픈소스 프로젝트 관리를 위해 가장 많이 사용되는 웹호스팅 서비스!!
- Git은 2005년 리눅스 제작자인 리누스 토발즈가 개발



- 가장 인기있는 오픈 소스 코드 저장소
- 깃허브 사용자는 2,800만 명 이상,
깃허브에 내에 저장된 소스코드 저장소는 약 5,700만 개
- 2018년 마이크로소프트가 인수

3. GitHub의 오픈소스 프로젝트

1.4 공유데이터와 오픈소스 (github)

	Contributors
1 <u>Microsoft/vscode</u>	19k
2 <u>facebook/react-native</u>	10k
3 <u>tensorflow/tensorflow</u>	9.3k
4 <u>angular/angular-cli</u>	8.8k
5 <u>MicrosoftDocs/azure-docs</u>	7.8k
6 <u>angular/angular</u>	7.6k
7 <u>ansible/ansible</u>	7.5k
8 <u>kubernetes/kubernetes</u>	6.5k
9 <u>npm/npm</u>	6.1k
10 <u>DefinitelyTyped/DefinitelyTyped</u>	6.0k

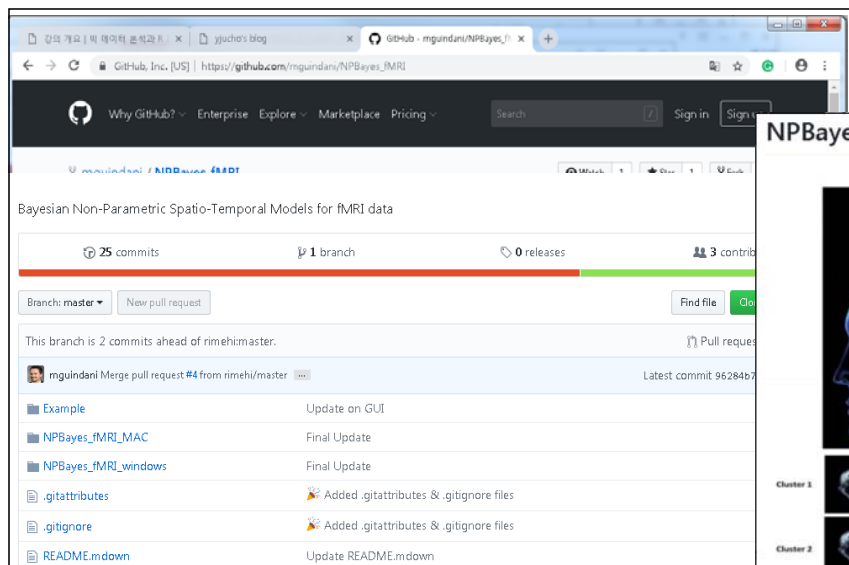
- VS코드는 **마이크로소프트**에서 개발해 오픈소스로 내놓은 코드 에디터
- 리액티브네이티브는 **페이스북**에서 공개한 크로스플랫폼 개발 프레임워크
- 텐서플로우는 **구글**에서 공개한 머신러닝 프레임워크

2017년 10월 1일 ~ 2018년 9월 30일 집계치

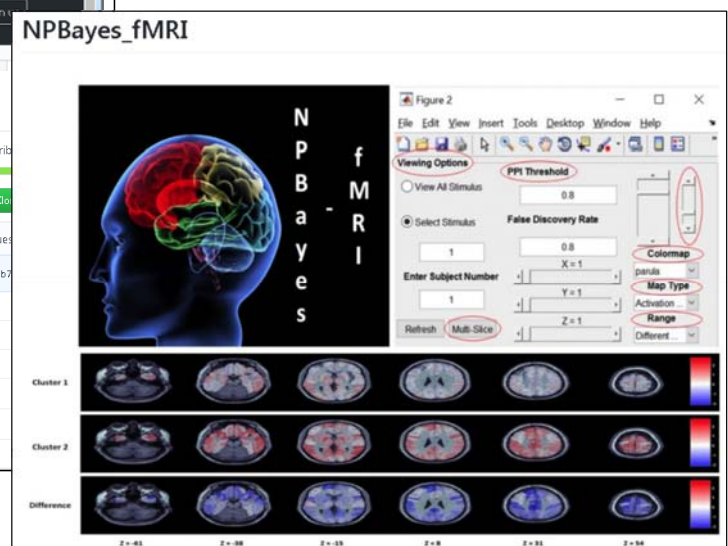
4. GitHub의 활용 : 오픈지식과 오픈코드

1.4 공유데이터와 오픈소스 (github)

- 예시 1 : fMRI데이터를 이용한 Naïve Bayes 기법적용



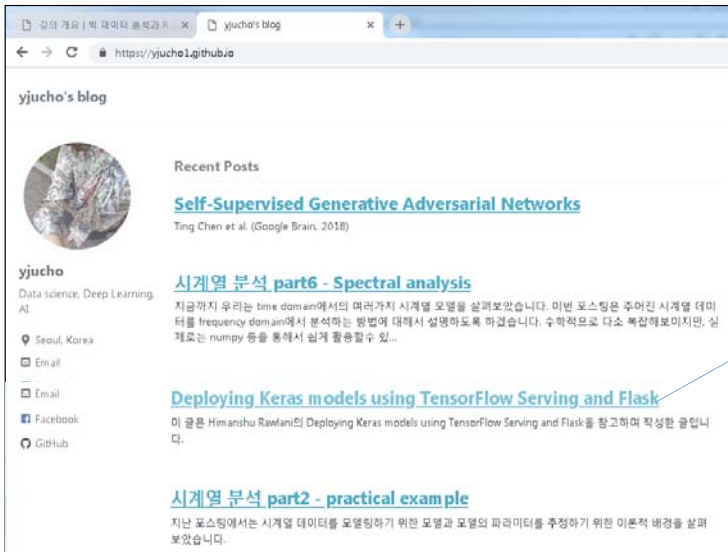
https://github.com/mguindani/NPBates_fmri



4. GitHub의 활용 : 오픈지식과 오픈코드

1.4 공유데이터와 오픈소스 (github)

• 예시2 : 시계열분석 정리내용? 혹은 Tensorflow를 이용한 Keras models?



https://yjucho1.github.io/

Deploying Keras models using TensorFlow Serving and Flask

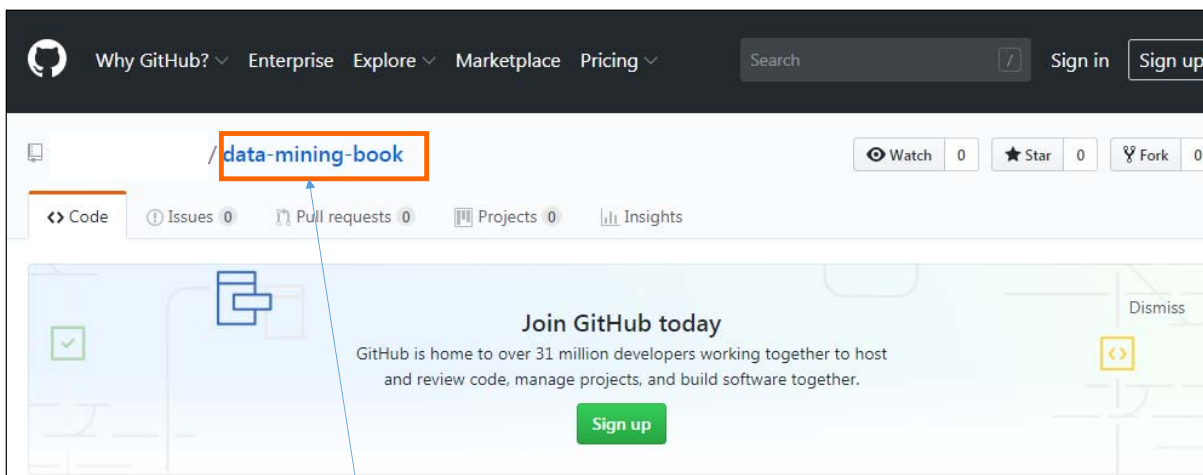
이 글은 Himanshu Rawlani의 Deploying Keras models using TensorFlow Serving and Flask을 참고하여 작성한 글입니다.

- 원글은 tensorflow serving 설치방법을 apt-get(리눅스 환경에서만 가능)을 이용하였으나, 본 글은 docker를 이용해 설치하는 방법을 이용해 MAC OSX에서 테스트한 내용을 추가하였습니다.
- 원글은 api서버로 Flask를 이용하였으나, 본 글은 django를 이용한 예제를 추가하였습니다. django APP의 전체코드는 [이 깃헙 레포지토리](#)를 참고하십시오.
- 추가된 부분은 Blockquotes 표시되어 있습니다.



4. GitHub의 활용 : 지식과 코드

1.4 공유데이터와 오픈소스 (github)



Github에 id를 생성하고 개별 프로젝트를 생성하여 공유하여 작업가능

- 공유데이터의 활용
- 공유데이터를 생성하고 제공
 - 오픈소스의 활용
 - 오픈소스의 개발과 서비스



2. 빅데이터 탐색의 첫걸음

	단위별 학습내용 (Week2)
wk2-1	데이터의 평균 (중심위치)
wk2-2	데이터의 분산 (산포정도)
wk2-3	데이터와 빅데이터
wk2-4	데이터 탐색의 첫걸음

Wk2-1 : 빅데이터 탐색의 첫걸음 - 데이터의 평균 (중심위치) -

1. 우리에게 이미 익숙한 평균

2.1 데이터의 평균 (중심위¹⁾)



중간고사가 끝난 후

이번 시험 잘 봤어?

나 **평균** 90점 받음!

1. 우리에게 이미 익숙한 평균

2.1 데이터의 평균 (중심위¹⁾)

• 다르게 말 할 수도 있다.



중간고사가 끝난 후

이번 시험 잘 봤어?

나 이번에
국어 80점, 수학 85점,
영어 80점, 사회 75점,
음악 90점, 체육 90점,
윤리 70점, 과학 90점,
국사 100점 받았어!

1. 우리에게 이미 익숙한 평균

2.1 데이터의 평균 (중심위)



평균 키가 큰 네덜란드 사람들

네덜란드 사람들은
정말 키가 커?

네덜란드 남자들은
평균 키가 183cm!

1. 우리에게 이미 익숙한 평균

2.1 데이터의 평균 (중심위)

• 누구도 이렇게 말하진 않는다.



평균 키가 큰 네덜란드 사람들

네덜란드 사람들은
정말 키가 커?

A는 키가 180cm,
B는 키가 183cm,
C는 키가 178cm,
...
ZZY는 키가 185cm,
ZZZ는 키가 183cm
이래!

평균은 데이터를
하나의 값으로 표현한
요약된 정보 (추정치) !

- 전과목의 평균 = (국어점수+수학점수+ ... +국사점수) / 9 (과목수)
- 네덜란드 성인남자 평균키 = 네덜란드 성인남자들 키 총합 / 조사대상자 수
- 평균 = 데이터 값의 총 합 / 데이터 개수

• n 개의 데이터 : x_1, x_2, \dots, x_n

평균 $\bar{x} = (x_1 + x_2 + \dots + x_n) / n = \sum_{i=1}^n x_i / n$

평균은
혼자 존재하는 개념이
아니다!

4. 평균을 다룰 때 주의할 점 – 평균과 표본선정

- 어떻게 표본선정을 하느냐에 따라 평균값에 영향을 미친다.

예) 대기업의 평균연봉을 알아보기 위해 표본 200명을 선정

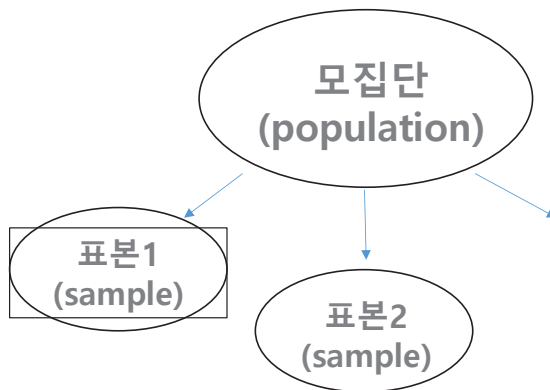


표본 A
20대~50대에서
각각 50명씩 선정

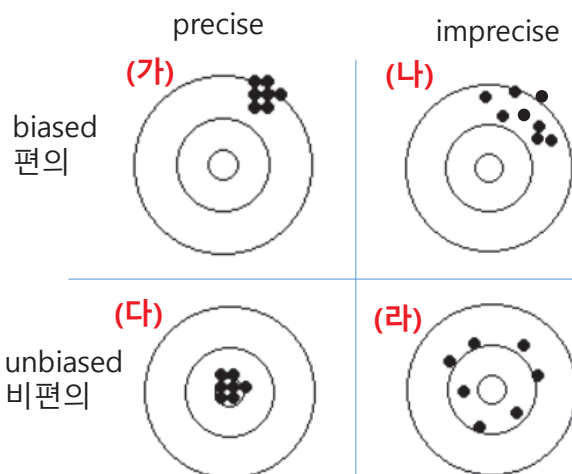


표본 B
50대에서 200명을 선정

조사된 평균값이
모집단을 대표하는 통계치라고 할 수 있나 ?



• 표본이 적합하게 추출되었는지 평가하는 방법 (평균을 예제로 하면)



1. 편익(Bias)가 적은가?

: 표본을 추출할 때 표본으로부터 얻어지는 통계치(표본 평균)의 기대값이 모수의 참값과 유사한가?

2. 정확도(Precision)가 높은가?

: 반복해서 표본을 추출할 때(반복 실험할때) 얼마나 유사한 값들이 나오는가?

4. 평균을 다룰 때 주의할 점 – 평균과 분산

2.1 데이터의 평균 (중심위¹⁾)

- 같은 평균이라도 분산이 다르면 데이터 특성은 다르다!



기업 A

평균 연봉 4,280만원
표준편차 2,399만원



기업 B

평균 연봉 4,280만원
표준편차 467만원

평균 연봉은 같지만
기업 A는 초봉이 낮고
임원 월급이 높겠구나!

4. 평균을 다룰 때 주의할 점

2.1 데이터의 평균 (중심위¹⁾)

평균값은
그 집단에서
가장 많이 존재하는 값이
아니다!

4. 평균을 다룰 때 주의할 점

2.1 데이터의 평균 (중심위¹⁾)

• 데이터 : 1, 2, 2, 7

• 평균 : 3



평균은 3이지만
3이 없다!

5. 데이터의 중심척도 요약

2.1 데이터의 평균 (중심위¹⁾)

• 평균(mean)은 표본이 적은경우 아주 큰 값이나 작은 값(outlier)에 민감한 추정치.
때로는 중앙값이 평균보다 더 적합한 중심척도인 경우도 있음

• 중앙값(median)

- n개의 관측치를 크기순으로 배열했을 때 중앙의 위치에 놓이게 되는 값
- 데이터의 수가 작고 이상치(outlier)가 있을 때 평균보다 더 정확한 모집단의 중심값이 됨

• 최빈값(mode)

- 전체 데이터 중 가장 빈도(frequency)가 높은 값
- 데이터의 수가 많아질수록 평균과 가까워짐



Wk2-2 : 빅데이터 탐색의 첫걸음

- 데이터의 분산 (산포정도) -

1. 어느 집단의 분산이 클까

- 데이터는 아는 만큼 보인다.
- 평균만 아는 사람 vs. 평균과 표준편차를 아는 사람



기업 A

평균 연봉 4,280만원



기업 B

평균 연봉 4,280만원

평균연봉이 같네!
직원 수도 50명으로 동일

1. 어느 집단의 분산이 클까

2.2 데이터의 분산 (산포정도)

- 데이터는 아는 만큼 보인다.
- 평균만 아는 사람 vs. 평균과 표준편차를 아는 사람



평균 연봉은 같지만,
기업 A는 초봉이 낮고
승진하면 월급이 높아지겠구나!

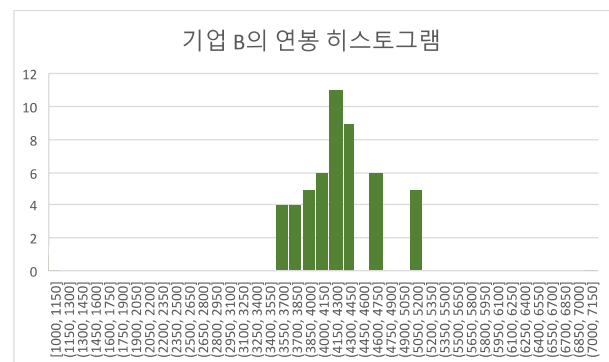
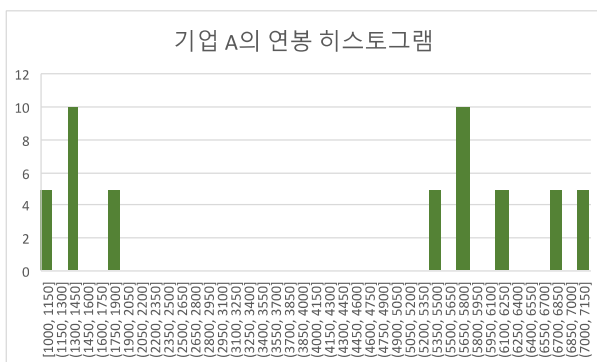
기업 A
평균 연봉 4,280만원
표준편차 2,399만원

기업 B
평균 연봉 4,280만원
표준편차 467만원

2. 그래프로 그려본 데이터의 산포

2.2 데이터의 분산 (산포정도)

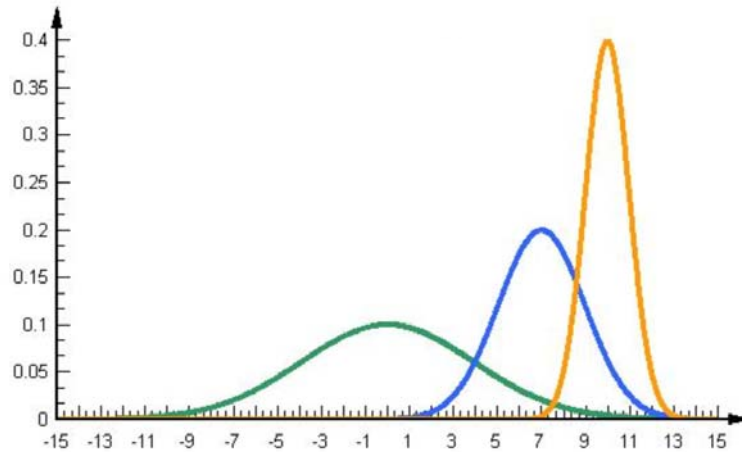
- 연봉의 히스토그램을 그려보자! (계급구간 너비=150만원)
- 기업 A는 양극단에 분포, 기업 B는 평균 중심에 많이 분포
→ 기업 A의 산포정도(분산)가 더 큼



2. 그래프로 그려본 데이터의 산포

2.2 데이터의 분산 (산포정도)

- 초록색의 분산이 가장 크고 노란색의 분산이 가장 작네!



3. 분산의 공식

2.2 데이터의 분산 (산포정도)

- 데이터의 산포정도가 크다
 - = 데이터가 중간에 몰려있지 않고 멀리 퍼져있다
 - = 데이터가 중심위치로부터 멀리 퍼져있다
 - = 데이터의 평균과 데이터들의 차이가 크다!

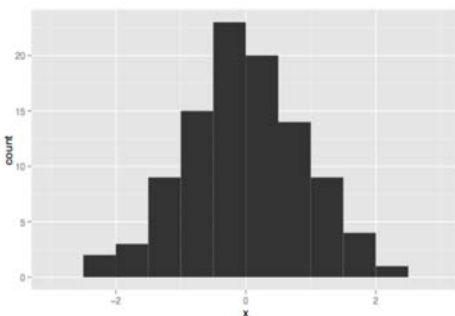
3. 분산의 공식

2.2 데이터의 분산 (산포정도)

- 데이터의 **평균**과 **데이터들의 거리의 합**으로 **분산**을 계산!
- 데이터 : x_1, x_2, \dots, x_n
- 평균 : \bar{x}
- 편차 : $(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$
- 편차들의 합 : $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = ?$

3. 분산의 공식

2.2 데이터의 분산 (산포정도)



- 편차들의 합 : $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = ?$
- 데이터가 평균으로부터 대칭적으로 존재할 경우 **편차들의 합이 0이 됨!**
→ 그래서 **편차를 제곱하여 더함!**
- **분산 = 편차들의 제곱합을 (n-1)*로 나눈다**
- **분산** :
$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

* (n-1)로 나누는 이유는 자유도와 관련, 평균값으로 표본평균을 사용하므로 1개의 자유도를 잃게 되어서 (n-1)로 나눈다.

3. 분산의 공식 - 표준편차

2.2 데이터의 분산 (산포정도)

중간고사 결과

평균 70 점

분산 225 점² (?)

표준편차 15 점

• (개별데이터값-평균값)의 차이를 제곱하여 더하였으므로 값이 커지고, 단위가 달라짐

→ 분산에 **제곱근**을 취하여 원래 단위로 복원

→ 이를 **표준편차**라고 부름

4. 분산의 의미

2.2 데이터의 분산 (산포정도)

분산은 데이터가
분포되어있는 정도

+

데이터에 대한 요약정보를 보완해줌!

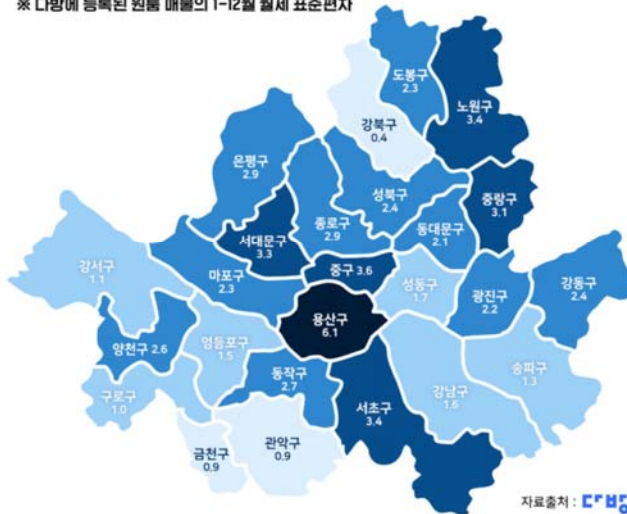
평균값만으로 데이터를 상상해보기 어려움!

4. 분산의 의미

2.2 데이터의 분산 (산포정도)

2018년 서울시 원룸 월세 변동성

※ 다방에 등록된 원룸 매물의 1-12월 월세 표준편차



용산구 월세

평균 52.6만원, 표준편차 6.14만원

vs.

강북구 월세

평균 35.2만원, 표준편차 0.37만원

4. 분산의 의미

2.2 데이터의 분산 (산포정도)



- 2018년 우리나라의 '행복지수' 평균은 157개국 중 57위로 비교적 높은 편
- 표준편차로 측정한 '행복 불평등도'는 157개국 중 96위로 행복의 격차가 매우 심각한 사회로 조사됨



Wk2-3 : 빅데이터 탐색의 첫걸음

- 데이터와 빅데이터 -

1. 데이터란 무엇인가

- 모든 숫자들은 데이터라고 할 수 있을까? No!

조사된 숫자 \neq 데이터

데이터란 **구조화**된 데이터!

1. 데이터란 무엇인가

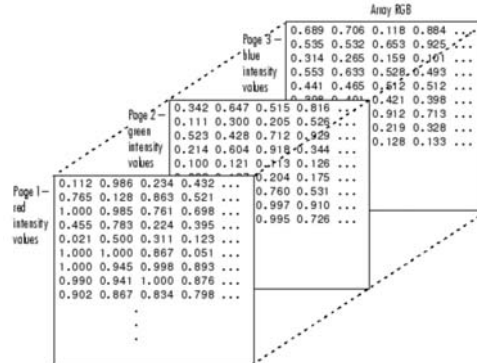
2.3 데이터와 빅데이터

	A	B	C	D	E	F	G	H	I
	License Number	Name	Address	City	ST	Zip	Date of Birth	Expiration Date	Gender
1	F298-6588	Anderson, Roger David	77 Sunset Strip	Miami	FL	33173	12/7/1952	12/7/2017	M
2	L781-9586	Balcock, George Hale	1000 College Blvd	Pensacola	FL	32504	5/17/1969	10/10/2017	M
3	T585-7121	Brewer, Larry Mitchell	4801 E Fowler Ave	Tampa	FL	33617	10/12/1956	10/12/2017	M
4	L998-5456	Castle, Frederick Evan	8581 Navarre Pkwy	Navarre	FL	32566	2/11/1980	3/30/2018	M
5	F742-5421	Cantrell, Carolyn Elise	1500 Miracle Strip Pkwy	Fort Walton Beach	FL	32548	7/11/1978	2/28/2018	F
6	T626-3357	Dixon, Cynthia Louise	366 13th St	Santa Rosa Beach	FL	32459	5/29/1952	5/15/2018	F
7	T929-8985	Evans, Susan Elaine	301 Hollywood Blvd E	Mary Esther	FL	32569	4/14/1953	6/11/2018	F
8	L303-2621	Garrett, Patrick Sean	44 Bayshore Point	Valparaiso	FL	32580	3/31/1972	3/31/2018	M
9	R881-9881	Hartley, Matthew Paul	500 Wondervood Dr	Jacksonville	FL	32233	8/21/1959	8/15/2018	M
10	R754-6523	Kensington, Carrie Ann	17000 Emerald Coast Pkwy	Destin	FL	32541	9/9/1979	10/19/2017	F
11	S755-0921	Lancouette, Phil	Margaritaville 500 Duval St	Key West	FL	33040	3/4/1960	3/4/2018	M
12	S181-1615	Mason, Daniel D	4607 State Park Lane	Panama City	FL	32408	1/22/1976	1/30/2018	M
13	L991-0220	Naylor, John T	900 N Birch Rd	Fort Lauderdale	FL	33304	6/30/1955	12/7/2017	M
14	R132-1895	Nicholas, Paul	6000 Universal Blvd	Orlando	FL	32819	2/3/1951	3/31/2018	M
15									
16									

• 데이터란 “구조화된 데이터”

- 다차원 배열(매트릭스)

- 각 열의 형식이 다른 표 or 스프레드시트



1. 데이터란 무엇인가

2.3 데이터와 빅데이터

time	channel1	channel2	channel3	channel4	channel5
1	1e-05	-2e-05	-1e-05	-3e-05	0
5	1e-05	-2e-05	-1e-05	-3e-05	0
6	-1e-05	1e-05	2e-05	0	1e-05
7	-1e-05	1e-05	2e-05	0	1e-05
8	-1e-05	1e-05	2e-05	0	1e-05
9	-1e-05	1e-05	2e-05	0	1e-05
10	-1e-05	1e-05	2e-05	0	1e-05
11	-1e-05	1e-05	2e-05	0	1e-05
12	-1e-05	1e-05	2e-05	0	1e-05
13	-1e-05	1e-05	2e-05	0	1e-05
14	-1e-05	1e-05	2e-05	0	1e-05
15	-1e-05	1e-05	2e-05	0	1e-05
16	-1e-05	1e-05	2e-05	0	1e-05
17	-1e-05	1e-05	2e-05	0	1e-05
18	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
19	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
20	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
21	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
22	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
23	-1e-05	-1e-05	-6e-05	-5e-05	-3e-05
24	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
25	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
26	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
27	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
28	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
29	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
30	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05
31	-3e-05	-1e-05	-4e-05	-1e-05	-1e-05

• 데이터란 “구조화된 데이터”

- 다차원 배열(매트릭스)

- 각 열의 형식이 다른 표 or 스프레드시트

- 탭이나 텍스트 파일 형식으로 저장

(* .txt, *.csv)

2. 데이터화 (Datafication)

2.3 데이터와 빅데이터

- 기계가 읽어들이 수 있는 모든것을 (숫자, 이미지, 텍스트) 데이터로 변환하는것
- 개인의 활동을 실시간으로 추적해 이를 예측분석이 가능한 수량화된 온라인 데이터로 변환하는 것을 의미한다 (Jose van Dijck, 2014)



Data
(숫자, 벡터로 변환)

2. 빅데이터란 무엇인가

2.3 데이터와 빅데이터

Volume(양) – 많은 양의 데이터

Velocity(속도) – 빠르게 생성

Variety(다양성) – 다양한 형태의 데이터



아마존닷컴의 추천 상품 표시

- 모든 고객들의 **구매 내역**을 데이터베이스에 기록
- 기록을 분석해 소비자의 **소비 취향과 관심사를 파악**
- 고객별로 추천 상품 (Recommendation)을 표시



영화 <머니볼>

- 머니볼 이론이란?
경기 데이터를 분석해
데이터를 기반으로 선수들을 배정
승률을 높인다는 게임 이론
- 최하위에 있던 팀을 4년 연속 포스트시즌에 진출시키고 메이저 리그 **최초로 20연승**이라는 신기록을 세움

- 데이터
- 데이터화
- 빅데이터
- 빅데이터의 활용



Wk2-4 : 빅데이터 탐색의 첫걸음

- 데이터 탐색의 첫걸음 -

데이터를 가지고 뭘 할 수 있을까?



1. 통계치로 인사이트를 얻는다

2.4 데이터 탐색의 첫걸음

• 방송사 공채에 합격하려면?



- 학점은 최소 3.6점 필요
- 토익점수 800점 이상 필요
- 수상내역을 특히 중요시함!

방송사 합격자 평균 분석 결과

1. 통계치로 인사이트를 얻는다

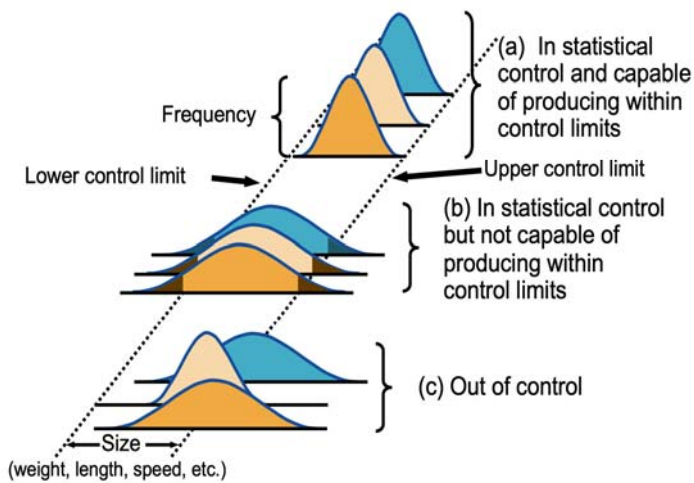
2.4 데이터 탐색의 첫걸음

• 사람들은 어떤 기업을 선호할까?



- 워라벨 열풍 속 복리후생이 중요해짐!

일하고 싶은 기업 기준 조사 결과



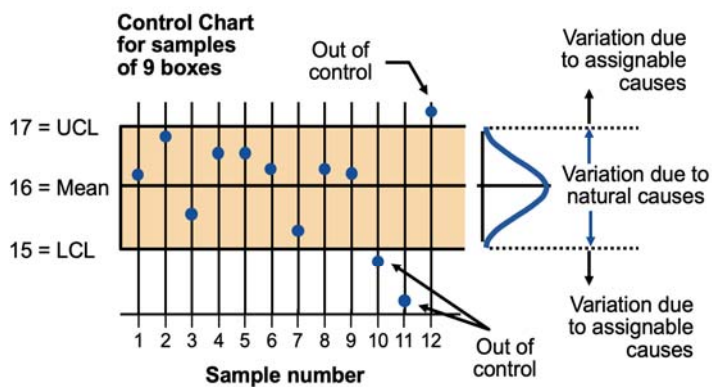
- 공정에 대한 **평균, 산포, 불량률**을 추정

→ 품질의 변동상황을 **관리도(Control Chart)**로 표현

→ 공정에 발생하는 이상요인을 빨리 탐지하여

수정조치를 취함으로써 **불량품 발생을 사전에 억제함**

관리도 차트



\bar{x} Chart

- 공정에서 정상범위 관리도 차트 : **정상일 때는 관리도 차트 내부에, 이상이 있을 때 벗어나는데**
-> **알람기능 (공정 관리자의 조정 필요)**
- **중심선, 관리상한선(UCL), 관리하한선(LCL)을 어떻게 설정할 것인가?**

2. 최적의 의사결정 – 통계적 품질관리

$$\text{관리상한선(UCL)} = \bar{\bar{x}} + z\sigma_{\bar{x}}$$

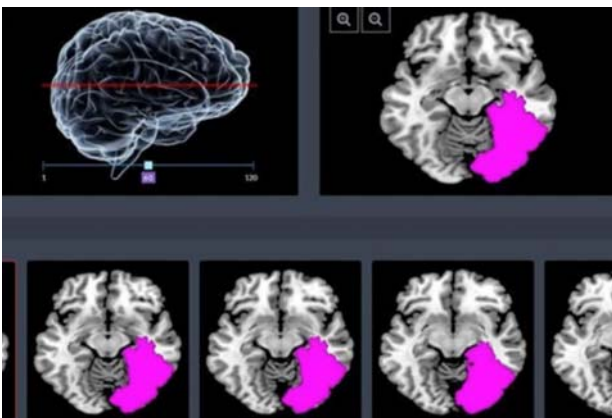
$$\text{관리하한선(LCL)} = \bar{\bar{x}} - z\sigma_{\bar{x}}$$

이렇게 중심선, UCL, LCL을
구하여 관리도를 만들 수 있음!

- $\bar{\bar{x}}$ = 표본평균들의 평균
- z = number of normal standard deviations (보통 3으로 설정)
- $\sigma_{\bar{x}}$ = 표본평균들의 표준편차

3. 데이터의 숨겨진 패턴을 분석 – 분류

- 이미지 분석을 통한 의료진단 및 헬스케어



암 여부 판단 및 수술 부위 판정

- 암과 정상인 뇌 영상을 숫자화(데이터화)함
 - 분류(암/정상)를 가장 잘 구분하는 변수를 찾고 범주 간의 차이를 가장 잘 표현하는 새로운 함수를 구함
 - 새로운 환자의 영상을 보고 어느 범주에 더 가까운지 판별하여 암 여부를 진다

3. 데이터의 숨겨진 패턴을 분석 - 분류

2.4 데이터 탐색의 첫걸음

- 각 영상은 p 개의 변수들로 이루어지며, 범주1(암) 또는 범주2(정상) 중 하나에 속함
- 변수들의 선형조합으로 새로운 변수 Z 를 형성 후 이를 바탕으로 분류규칙을 만듦

$$Z = w_1X_1 + w_2X_2 + \dots + w_pX_p = w^T x$$

- 두 범주가 잘 분류된다는 것은 **두 범주가 겹치지 않으면서** 두 범주의 **중심위치가 가능한 멀게**

→ 범주 간 Z 의 $\lambda = \frac{\text{범주 간 } Z \text{의 평균차이}}{Z \text{의 분산}}$ 값이 **최대화되는 w 값**을 찾는 것이 목적

3. 데이터의 숨겨진 패턴을 분석 - 분류

2.4 데이터 탐색의 첫걸음

- $\lambda = \frac{\text{범주 간 } Z \text{의 평균차이}}{Z \text{의 분산}}$ 는 $\lambda = \frac{w^T(\mu_1 - \mu_2)}{w^T \Sigma w}$, 즉 평균과 분산으로 나타낼 수 있음

- 범주 구분을 위해 λ 의 최대값을 구하기 위해 w 에 대해 미분하면 $w = \Sigma^{-1}(\mu_1 - \mu_2)$

- 새로운 데이터의 범주를 분류하기 위해 각 범주의 **표본평균과 판별함수값 Z 와의 차이**를 산출 후
그 차이가 가장 작은 범주에 분류함

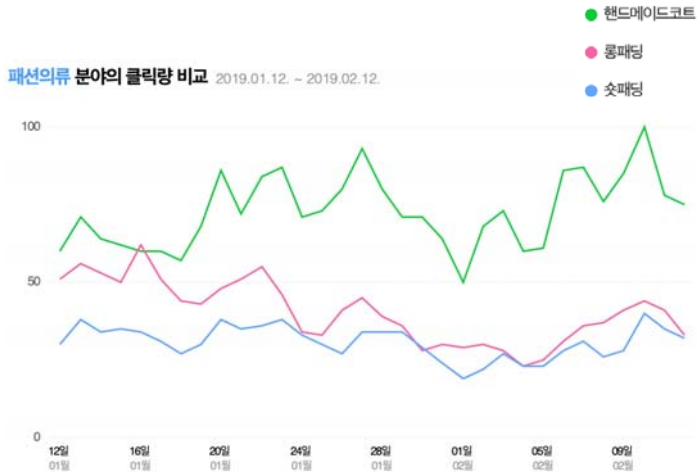
$$|\hat{w}^T(x - \bar{x}^{(1)})| \leq |\hat{w}^T(x - \bar{x}^{(2)})| \text{ 이면, } x \text{를 범주1로 분류}$$

$$|\hat{w}^T(x - \bar{x}^{(1)})| > |\hat{w}^T(x - \bar{x}^{(2)})| \text{ 이면, } x \text{를 범주2로 분류}$$

4. 웹 마이닝을 통한 트렌드 분석

2.4 데이터 탐색의 첫걸음

• 지난 1년간 검색어 트렌드 비교



• 지난 1년간의 트렌드 분석

핸드메이드코트가 가장 인기,
롬패딩의 인기는 작년 겨울에 비해 감소,
숏패딩은 전반적으로 인기가 가장 낮음

• 앞으로의 트렌드 예측

지난 1년의 트렌드와 비슷할 것이라는 가정
하에 앞으로의 트렌드를 예측해볼 수 있음

4. 웹 마이닝을 통한 트렌드 분석!

2.4 데이터 탐색의 첫걸음

• Weighted Moving Average를 통한 트렌드 파악

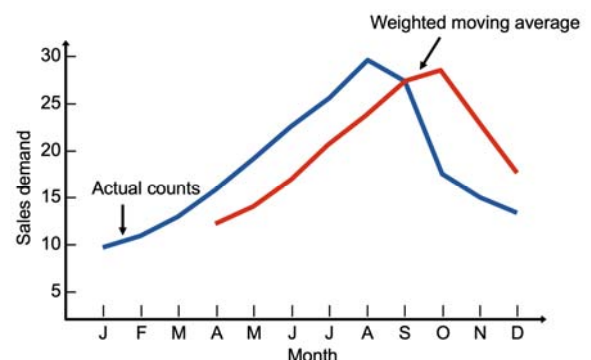
→ 과거 트렌드를 반영하되 먼 과거의 데이터보다 가까운 시점의 데이터를 더 중요시함!

$$WMA = \frac{\sum ((\text{Weight for period } n) (\text{Demand in period } n))}{\sum \text{Weights}}$$

MONTH	ACTUAL COUNTS	3-MONTH WEIGHTED MOVING AVERAGE
January	10	
February	12	
March	13	
April	16	$[(3 \times 13) + (2 \times 12) + (1 \times 10)] / 6 = 12 \frac{1}{6}$
May	19	
June		
July		
August		
September		
October		
November		
December		

WEIGHTS APPLIED	PERIOD
3	Last month
2	Two months ago
1	Three months ago
6	Sum of the weights

Forecast for this month = $3 \times \text{Sales last mo.} + 2 \times \text{Sales 2 mos. ago} + 1 \times \text{Sales 3 mos. ago}$
Sum of the weights



- 제대로 된 데이터가 있다면
 - 통계치를 도출하여 대상에 대한 인사이트를 얻을 수 있다.
 - 미래 현상을 예측하여 의사결정을 할 수 있다.
 - 통계적 개념을 바탕으로 유용한 차트를 만들어낼 수 있다.
 - 분류 분석, 트렌드 분석 등에 활용할 수 있다.



3. 데이터시각화와 통계적 해석

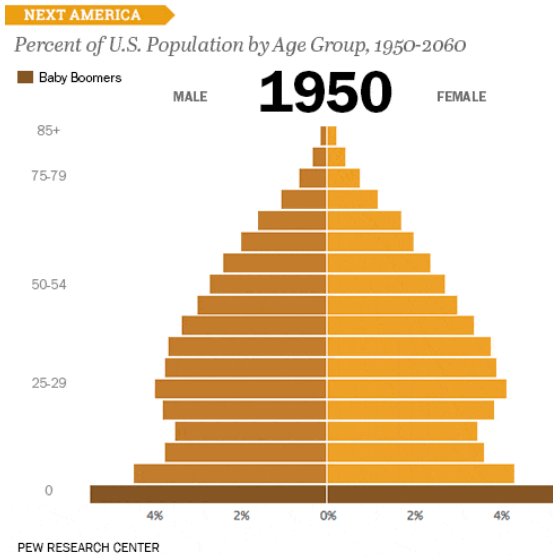
	단위별 학습내용 (Week3)
wk3-1	데이터 시각화
wk3-2	그래프의 유용성과 오류
wk3-3	상자그림이 주는 정보와 해석
wk3-4	산점도와 상관관계 - 트렌드 분석

Wk3-1 : 데이터 시각화와 통계적 해석

- 데이터 시각화 -

1. 데이터 시각화란?

3.1 데이터 시각화



- 데이터 분석 결과를
쉽게 이해할 수 있도록 보여주는 것!
- 그래프, 도표, 이미지, 단어 구름 등을
통해 한 눈에 이해할 수 있도록 하는 것!

1. 데이터 시각화란?

3.1 데이터 시각화

수집
(Data gathering)

정제
(Data processing)

시각화
(Data
visualization)

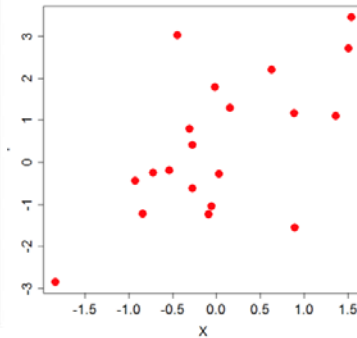
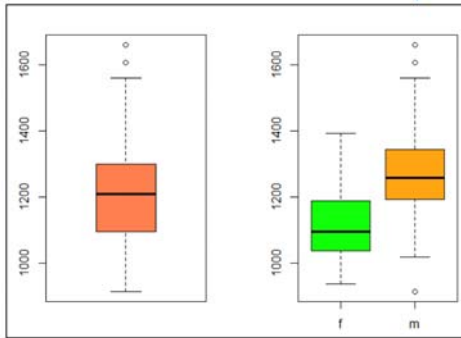
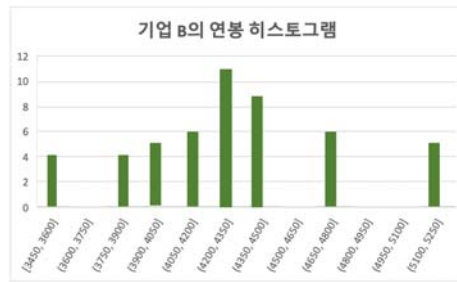
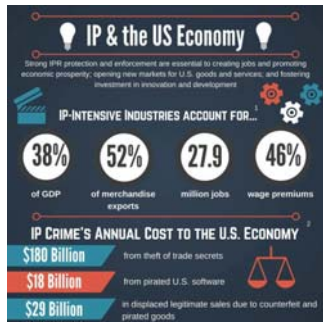
예측모형/분석
(Data analysis)

데이터 분석단계



1. 데이터 시각화란?

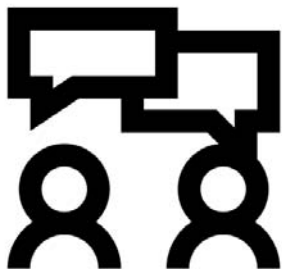
3.1 데이터 시각화



- 인포그래픽(Infographic)
- 히스토그램(Histogram)
- 상자그림(Box plot)
- 산점도(Scatter plot)

2. 데이터 시각화를 하는 이유

3.1 데이터 시각화



Communication



Discovery



Insight

3. 효과적인 데이터 시각화의 조건

3.1 데이터 시각화

- 어떤 메시지를 전달할 것인지 결정 (What)
- 핵심 내용을 제외한 나머지는 생략 (What)
- 최선의 표현 방법을 선택 (How)
- 단순, 명료하게 디자인 (How)
- 데이터를 토대로 어떤 의사결정을 해야 하는지에 대해 설명 (Why)

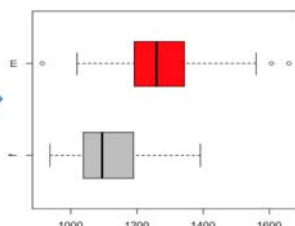
4. 데이터 시각화 도구

3.1 데이터 시각화

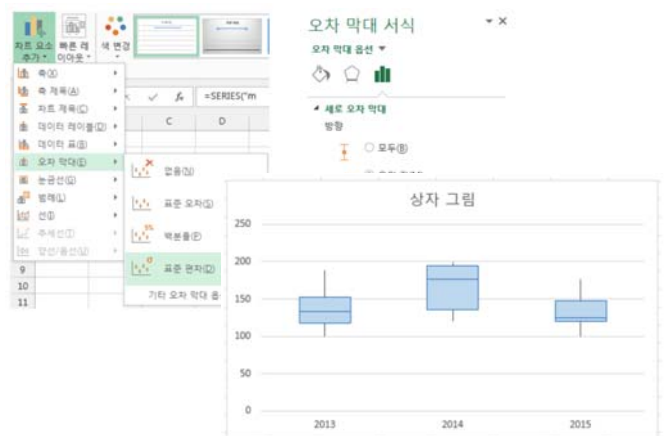
- R, Python, 엑셀 등

```
# 2-3 horizontal boxplot  
par(mfrow=c(1,1))  
boxplot(brain$wt~brain$sex, boxwex=0.5, horizontal=TRUE, col = c("grey", "red"))
```

수평으로 상자그림을 그릴수 있음



R로 상자그림을 그리는 방법



엑셀로 상자그림을 그리는 방법



금융 데이터 모니터링을 위한 대시보드 화면

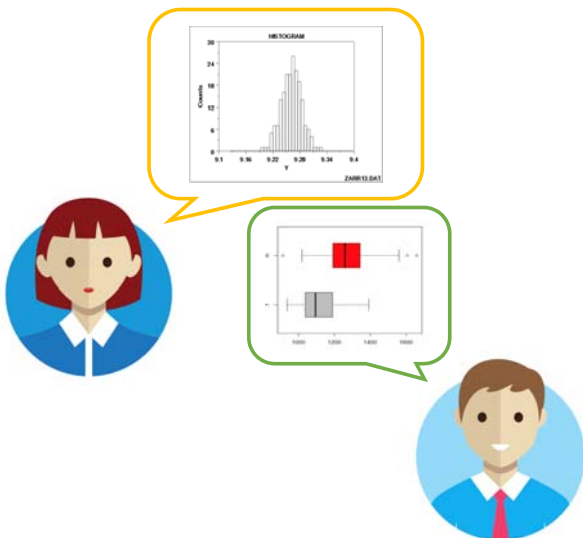
- 데이터 분석에 대한 전문성이 없어도 누구나 데이터를 직관적으로 이해
- 효과적으로 인사이트 도출
- 기업의 의사결정에서 근거자료



Wk3-2 : 데이터 시각화와 통계적 해석

- 그래프의 유용성과 오류 -

1. 그래프의 유용성

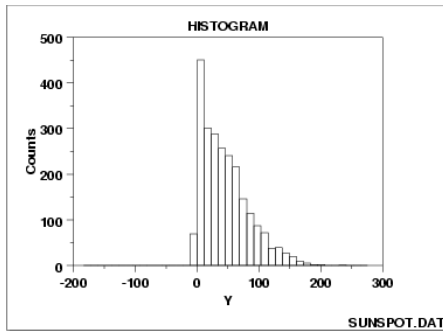


- 그래프는 **데이터 시각화**의 일종
- 그래프의 올바른 해석은
데이터사이언티스들의 **필수 능력**이자
그들의 **커뮤니케이션 도구**이다.

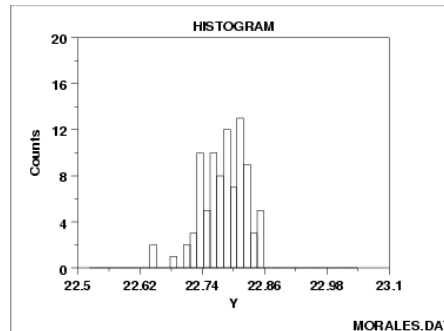
1. 그래프의 유용성

3.2 그래프의 유용성과 오류

• 히스토그램으로 보면



Right skewed



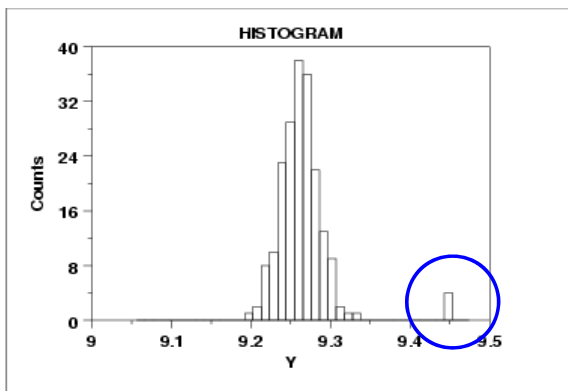
Left skewed

- 같은 분산이라도 데이터의 분포를 더 잘 파악할 수 있음

1. 그래프의 유용성

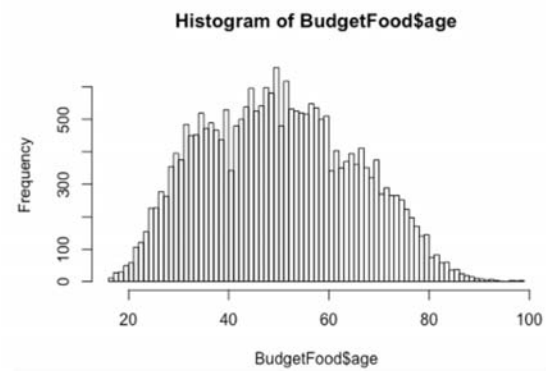
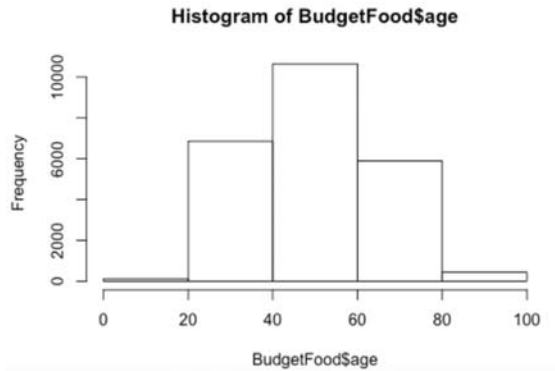
3.2 그래프의 유용성과 오류

• 히스토그램으로 보면



- 이상치(Outlier)의 존재도 파악 가능

2. 그래프의 왜곡



계급 구간을 어떻게 설정하는지에 따라
히스토그램 그래프가 완전히 달라짐!

2. 어떻게 시각화를 해야 정확한 정보를 제공할 수 있다?

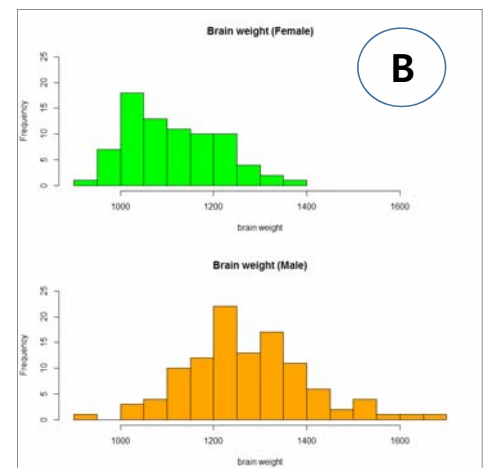
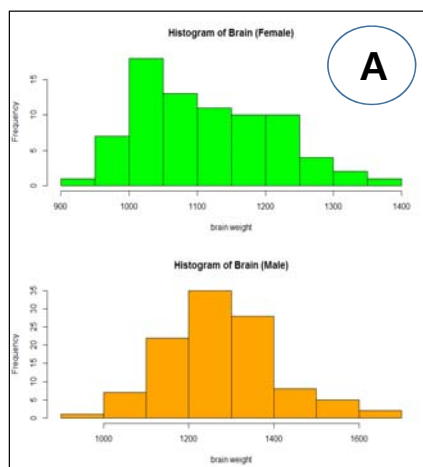
- 어느 그래픽이 더 나을까?
- 한눈에 어느 그룹(female/male)의 뇌의 무게가 무거운지 금방 알수 있나?

A : 동일하지 않은 범위

B : 동일한 범위로 시각화



B 그래픽은 한눈에 봐도 금방
정보(결과)를 읽을수 있음!!



3. R코드 예제

3.2 그래프의 유용성과 오류

• R코드 (hist3_2.r) – 데이터 brain.csv 불러들이기, 현재작업폴더 지정하기

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
hist3_2.R
1 # hist3_2.r
2 # histogram by group
3
4 # set working directory
5 # change working directory
6 setwd("D:/tempstore/stat")
7
8 brain<-read.csv("brain.csv")
9 head(brain)
10
11 attach(brain)
12
13 # histogram for all data
14 hist(wt, col="lightblue")
```

```
# histogram for female and male
# 2*1 multiple plot
par(mfrow=c(2,1))
brainf<-subset(brain,brain$sex=='f')
hist(brainf$wt, breaks = 12,col = "green",cex=0.7, main="Histogram of Brain (Female)",
# subset with male
brainm<-subset(brain,brain$sex=='m')
hist(brainm$wt, breaks = 12,col = "orange", main="Histogram of Brain (Male)", xlab="br
# Histogram with same scale and range
# 2*1 multiple plot
par(mfrow=c(2,1))
# histogram with same scale
hist(brainf$wt, breaks = 12,col = "green", xlim=c(900,1700),ylim=c(0,25), main="Brain
hist(brainm$wt, breaks = 12,col = "orange", xlim=c(900,1700), ylim=c(0,25),main="Brain
```

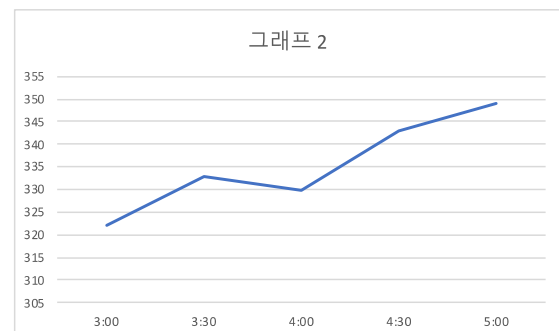
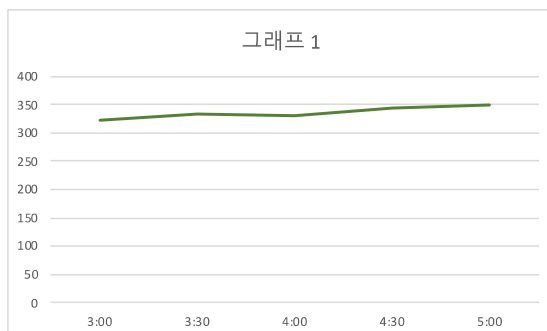
par(mfrow=c(2,1)) : 그래프화면의 분할을 행(row)는 2행으로, 열은 1열 의미

x축의 범위 : xlim=c(0,100), y축의 범위 : ylim(0,100)

[포스텍MOOC] <https://pabi.smartlearn.kr/>에서 '빅데이터와 R프로그래밍' 수강가능
- 강좌미리보기 week1_1(설치), week3_1(데이터불러들이기), week4_1(그래픽 I) 참조

4. 그래프의 왜곡

3.2 그래프의 유용성과 오류

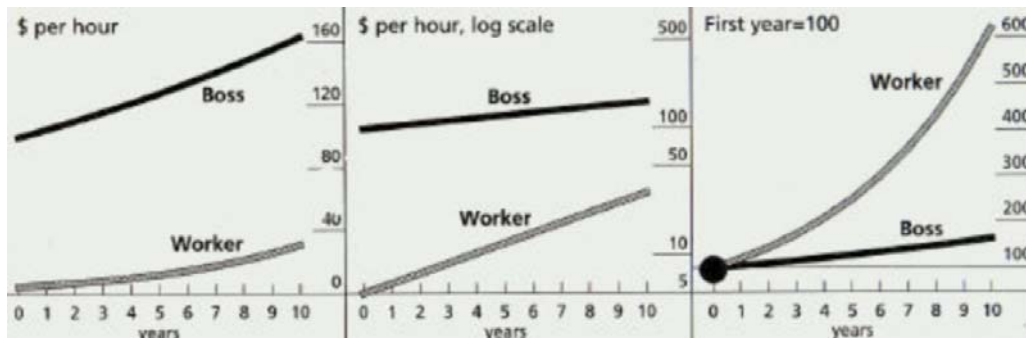


Y 축의 시작, 끝 값 설정에 따라
차이가 있어보이기도 하고 없어보이기도 함!

4. 그래프의 왜곡

3.2 그래프의 유용성과 오류

경영진과 근로자의 임금 증가율



실제 시간당 임금증가

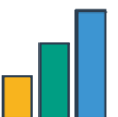
Y 축의 설정에 따라
그래프가 전혀 다르게 나타남!

5. 데이터 시각화의 주의할 점

3.2 그래프의 유용성과 오류

• 그래프의 목적은 데이터를 분명하게 표현하는데 있다.

1. 그래프를 작성할 때는 축의 범위, 간격 등을 잘 정해야 한다.
2. 그래프를 보는 사람의 수준을 고려해야 한다.
3. 그래프 종류별 장점과 단점을 정확히 파악하고 사용해야 한다.



Wk3-3 : 데이터 시각화와 통계적 해석

- 상자그림이 주는 정보와 해석 -

1. 왜 상자그림이 필요한가?

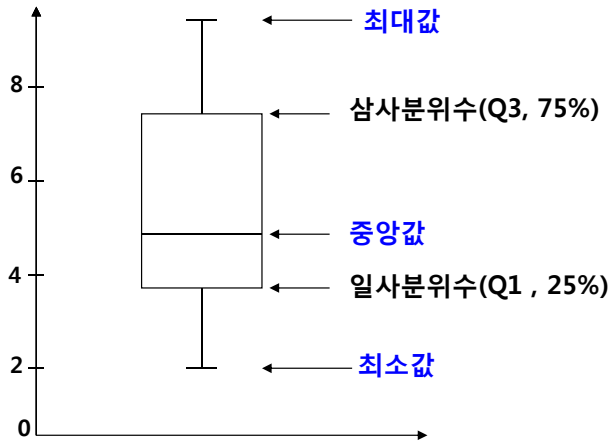
평균과 분산만으로는
부족하다!

기술통계치만으로는 데이터
에 대해 잘 알 수가 없다!

- 평균, 분산을 안다고 데이터가 어느 쪽에 더 많이 분포하는지 알 수 없다
- 이상치가 존재하는지 알 수 없다
- 데이터의 분포 범위(최대값, 최소값)을 한눈에 알기 어렵다

2. 상자그림이 주는 정보

3.3 상자그림이 주는 정보와 해석



- 상자그림은 한눈에 5가지 정보를 제공
(중앙값, 일사분위수, 삼사분위수, 최대값, 최소값)

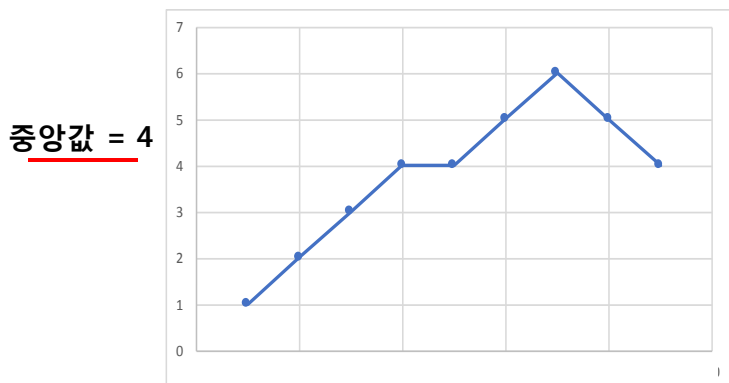


데이터 분포의 **대칭성, 치우침, 이상치**를
쉽게 파악할 수 있음!

3. 상자그림을 그리는 방법

3.3 상자그림이 주는 정보와 해석

(1) 데이터의 **중앙값(median)**을 찾는다.



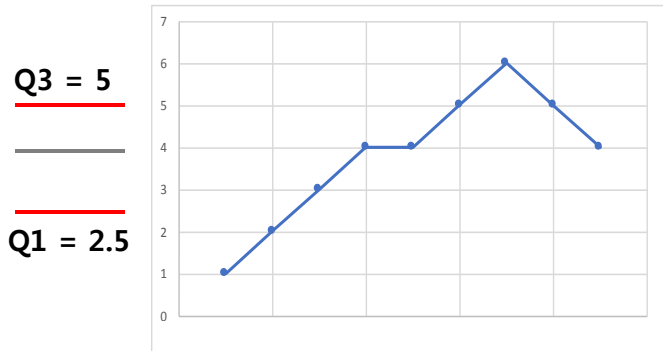
• 중앙값이란?

- n개의 관측치를 오름차순으로 배열했을 때 중앙의 위치에 놓이게 되는 값
- 데이터의 수가 작고 이상치가 있을 때 평균보다 더 정확한 모집단의 중심값이 됨

3. 상자그림을 그리는 방법

3.3 상자그림이 주는 정보와 해석

(2) 일사분위수(Q1)과 삼사분위수(Q3)을 찾는다.



- 일사분위수(Q1)

데이터를 크기순서로 배열했을때 25% 지점의 값

- 삼사분위수(Q3)

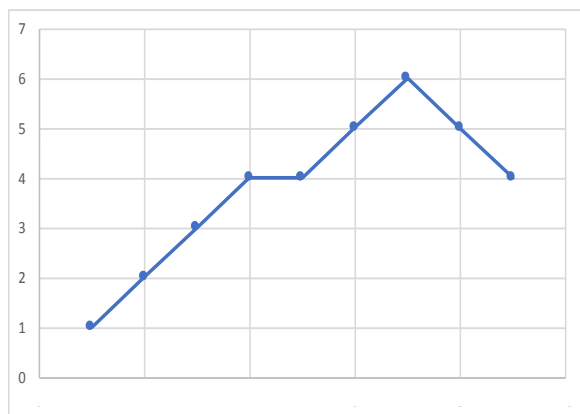
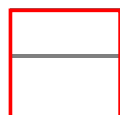
데이터를 크기순서로 배열했을때 75% 지점의 값

3. 상자그림을 그리는 방법

3.3 상자그림이 주는 정보와 해석

(3) 일사분위수 ~ 삼사분위수를 **상자**로 그린다! (사분위범위)

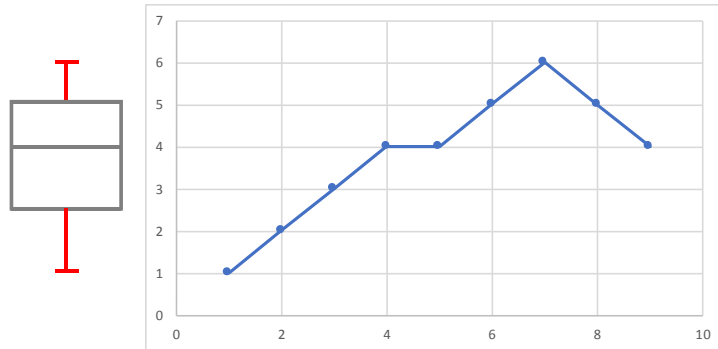
상자(Box)



3. 상자그림을 그리는 방법

3.3 상자그림이 주는 정보와 해석

(4) 최소값~일사분위수, 삼사분위수~최대값을 그린다!



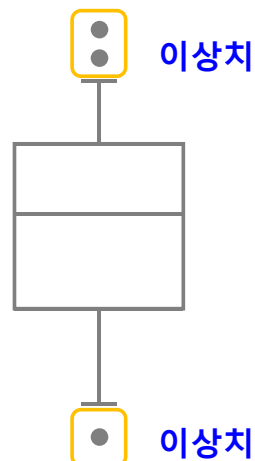
4. 상자그림의 해석

3.3 상자그림이 주는 정보와 해석

(5) 이상치 표시

일사분위로부터 $-(1.5) \times$ 사분위범위를 넘는 관측치는 이상치로 표시한다.

삼사분위로부터 $+(1.5) \times$ 사분위범위를 넘는 관측치는 이상치로 표시한다.



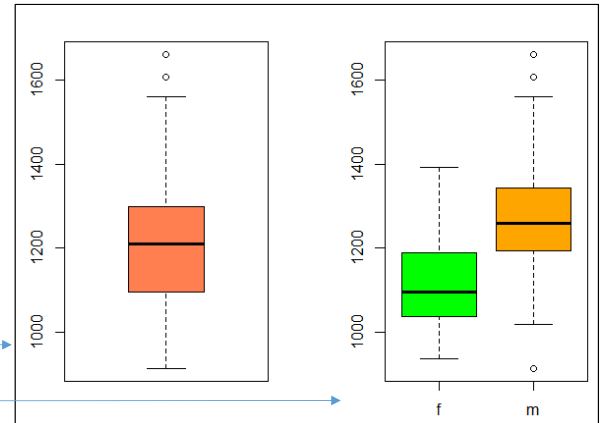
5. R 코드 (Boxplot)

3.3 상자그림이 주는 정보와 해석

- 상자그림 : `boxplot(변수이름, col=c("colname"))`
- 그룹별 상자그림 : `boxplot(wt~sex, col=c("green", "orange"))`

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
hist3_2.R x boxplot3_3.R x
1 # boxplot3_3.r : Boxplot
2
3 # set working directory
4 # change working directory
5 setwd("D:/tempstore/stat")
6
7 brain<-read.csv("brain.csv")
8 head(brain)
9
10 attach(brain)
11 |
12 # 2. boxplot
13 par(mfrow=c(1,2))
14 # 2-1 boxplot for all data
15 boxplot(wt, col=c("coral"))
16 # 2-2 boxplot by group variable (female, male)
17 boxplot(wt~sex, col = c("green", "orange"))
```

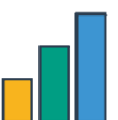
`par(mfrow=c(1,2))` : 그래프화면의 분할을 행 (row)는 1행으로, 열은 2열로 하라는 의미



6. 요약

3.3 상자그림이 주는 정보와 해석

- 상자그림은 다섯가지 숫자로 데이터를 요약한 그래프이다.
- 가운데 있는 상자는 **Q1** 에서 **Q3** 까지 그린다.
- 상자 안에 있는 선은 **중앙값**을 나타낸다.
- 상자 밖의 선은 **최대값**과 **최소값**까지 이어진다.
- 상자와 수염 밖의 데이터는 이상치이다

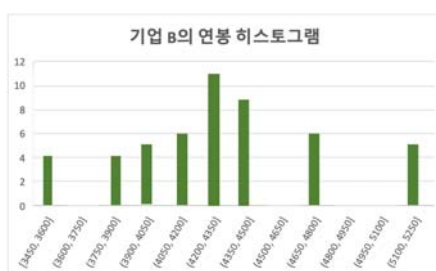


Wk3-4 : 데이터 시각화와 통계적 해석

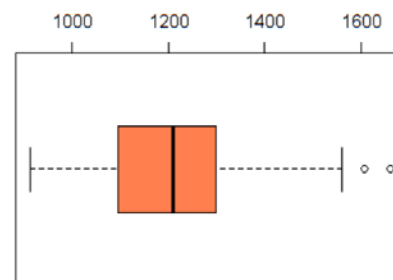
- 산점도와 상관관계 : 트렌드 분석 -

1. 왜 산점도가 필요한가?

- 지금까지 히스토그램, 상자그림을 통해 변수 1개의 데이터 분포를 살펴봄
- 그렇다면 **두 변수 사이의 관계**는 어떻게 알 수 있을까?



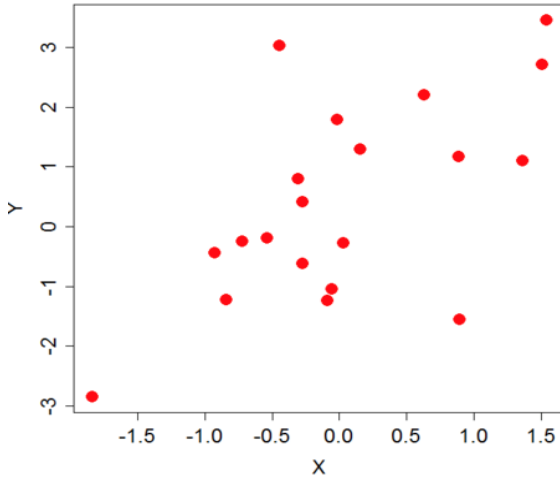
히스토그램



상자그림

1. 왜 산점도가 필요한가?

3.4 산점도와 상관관계 : 트렌드 분석



- 변수 간 관계의 방향, 트렌드, 강도를

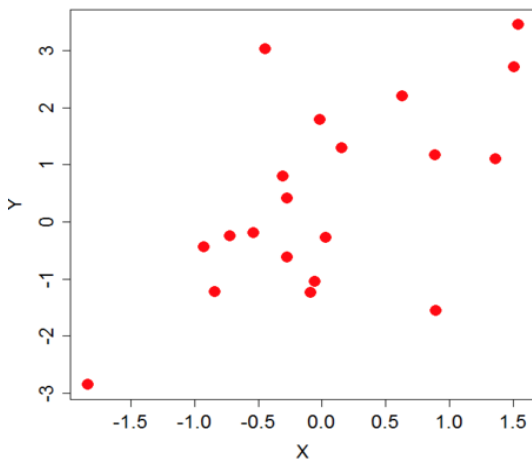
살펴볼 수 있는 그래프가

산점도(Scatter plot!)

1. 왜 산점도가 필요한가?

3.4 산점도와 상관관계 : 트렌드 분석

- 산점도의 x축과 y축은 독립변수와 종속변수로 이루어짐



- 독립변수 (Independent/Explanatory variable)

: 원인의 역할을 하는 변수, X

- 종속변수 (Dependent/Response variable)

: 결과를 관측하는 변수, Y

예시 : CO₂ Level - 자동차 가솔린량,

학점 - 공부한 시간, 수명 - 흡연 여부

1. 왜 산점도가 필요한가?

3.4 산점도와 상관관계 : 트렌드 분석

- 산점도로부터 알 수 있는 3가지

1. **트렌드** : linear, curved, clusters, no pattern
2. **방향** : positive, negative, no direction
3. **강도** : how closely the points fit the trend

2. 산점도의 해석 - 방향

3.4 산점도와 상관관계 : 트렌드 분석

- 두 변수 X와 Y가

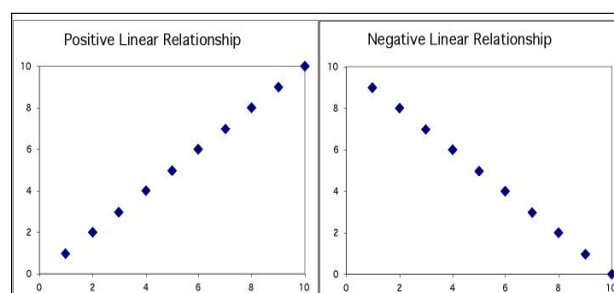
X값이 클 때 Y값도 큰 경향이 있고, X값이 작을 때 Y값도 작은 경향이 있다면?

→ 양의 상관관계에 있음! (Positively associated!)

- 두 변수 X와 Y가

X값이 클 때 Y값이 작은 경향이 있고, X값이 작을 때 Y값이 큰 경향이 있다면?

→ 음의 상관관계에 있음! (Negatively associated!)



2. 산점도의 해석 - 상관계수

- 상관관계의 **강도**를 나타내는 것이 **상관계수(Correlation, r)**

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

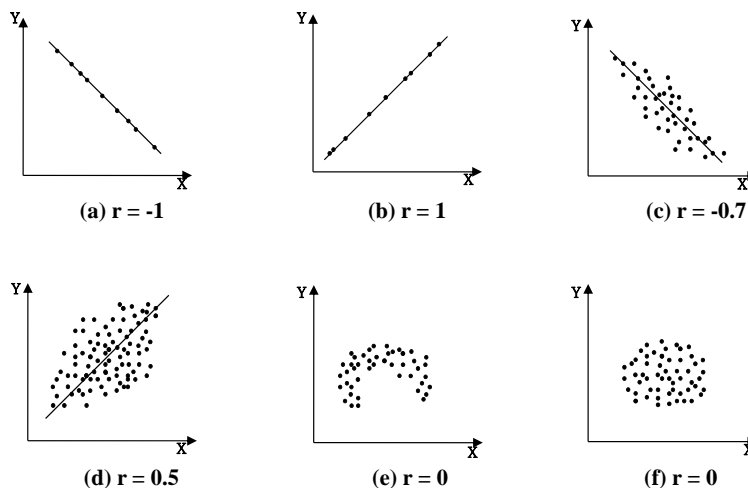
r 은 **-1** 부터 **+1**까지 존재하며,

+1에 가까울수록 강한 양의 상관관계, **-1**에 가까울수록 강한 음의 관계에 있음!

0은 가장 약한 상관관계를 의미함 (상관관계가 없음)

2. 산점도의 해석 - 상관계수

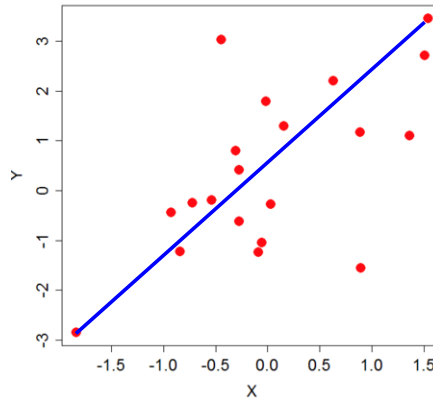
- 산점도에서 상관관계란 **선형적인(linear)** 상관관계만 의미함!



2. 산점도의 해석 - 상관계수

3.4 산점도와 상관관계 : 트렌드 분석

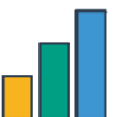
- 산점도에서 상관관계란 **선형적인(linear)** 상관관계만 의미함!
 - 두 변수를 **한 직선**으로 표현할 수 있지 않을까?
 - 이 직선을 토대로 **임의의 X값에 대한 Y값을 예측**할 수 있지 않을까?



3. 요약

3.4 산점도와 상관관계 : 트렌드 분석

- 산점도는 두 변수 간 관계의 방향, 형태, 강도를 살펴볼 수 있는 그래프이다.
- 상관계수(r)는 두 변수 간 선형적인 상관관계의 강도를 나타낸다.
- 산점도에서 선형모형(선형함수식)을 구현할수 있다.



4. 빅데이터분석에서의 확률과 분포

	단위별 학습내용 (Week4)
wk4-1	확률의 기초개념
wk4-2	조건부확률과 베이지확률
wk4-3	정규분포(연속형)과 포아송분포(이산형)
wk4-4	데이터에서 출발하는 확률과 분포

Wk4-1 : 빅데이터분석에서의 확률과 분포 - 확률의 기초개념 -

1. 통계에서 확률 개념은 왜 필요한가?

통계(Statistics)

≠

확률(Probabilities)

- **통계**란 데이터를 수집, 처리, 분석, 활용하는 지식
→ **실제 얻어진 데이터를 바탕으로 정보를 도출**
- **확률**이란 어떤 특정한 사건이 일어날 가능성을 0 과 1 사이의 값으로 나타낸 것
→ **관측하기 전에 있어서 가능성을 논하는 것**

1. 통계에서 확률 개념은 왜 필요한가?



- 현실 세계는 매우 랜덤하다.
미리 그 결과를 알 수 없다.
→ **단기적으로** 어떠한 사건이 일어날 비율은 매우 **랜덤함!**

1. 통계에서 확률 개념은 왜 필요한가?



- 하지만 장기적으로 어떤 사건이 일어날 가능성은 확률적으로 예측 가능하다!
- 사건 하나하나에 대해서 미리 아는 것은 불가능하더라도 확률적 모형을 통해 많은 시행의 결과를 예측할 수 있음!



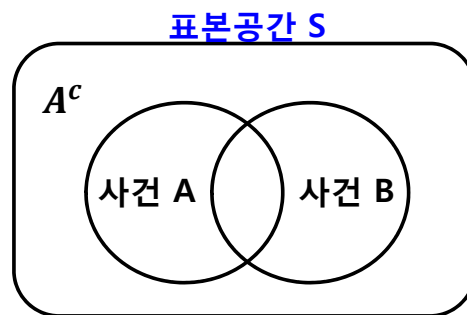
2. 통계에서 필요한 확률 – 확률, 사건, 표본공간

- 확률 : 어떤 특정한 사건이 일어날 가능성을 0과 1 사이의 값으로 나타낸 것
- 사건 : 표본공간에서 관심의 대상인 부분집합
- 표본공간 : 확률실험의 모든 가능한 결과의 집합

$$P(A) = \frac{\text{사건 } A \text{가 일어나는 경우의 수}}{\text{모든 가능한 결과의 수}}$$

2. 통계에서 필요한 확률 – 확률, 사건, 표본공간

- **합집합사건**: 사건 A 또는 사건 B가 일어날 때, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **교집합사건**: 사건 A와 사건 B가 동시에 일어날 때, $P(A \cap B)$
- **여집합사건**: 표본공간 S에서 사건 A가 일어나지 않을 때, $P(A^c) = 1 - P(A)$
- **배반사건**: 교집합사건이 공사건일 때, 사건 A와 B가 서로 배반(mutually exclusive)



2. 통계에서 필요한 확률 – 예시

(예제 1)

하나의 주사위를 던지는 실험에서 표본공간은 $S = \{1, 2, 3, 4, 5, 6\}$ 이다.

A를 홀수가 나올 사건, B를 4이하의 수가 나올 사건이라 하면,

A와 B의 합집합사건과 그 확률, A와 B의 교집합사건과 그 확률은 무엇인가?

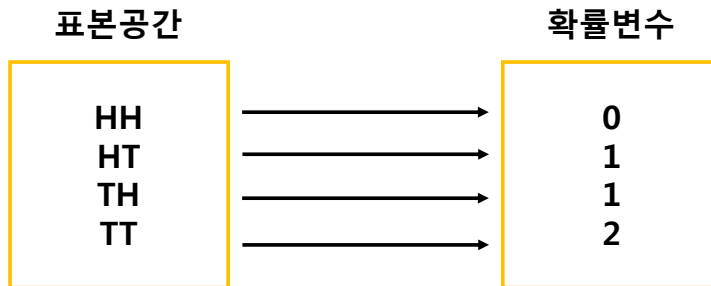
• 정답) $A = \{1, 3, 5\}$, $B = \{1, 2, 3, 4\}$

$A \cup B = \{1, 2, 3, 4, 5\}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 5/6$

$A \cap B = \{1, 3\}$, $P(A \cap B) = 1/3$

3. 통계에서 필요한 확률 – 확률변수와 기대값

- **확률변수** : 확률 실험으로부터 나타난 **결과에 실수를 할당한 함수**



- **기대값** : **확률변수의 중심척도**로, 어떤 랜덤한 상황에서 수치로 나타난 결과가

A_1, A_2, \dots, A_K 이고 각 결과의 확률이 P_1, P_2, \dots, P_K 이면 기대값은 각 결과에

확률을 곱하여 전부 합한 것

$$\text{기대값} = A_1 P_1 + A_2 P_2 + \dots + A_K P_K$$

3. 통계에서 필요한 확률 – 예시

(예제 2)

앞면이 나올 확률이 $\frac{1}{2}$ 인 동전을 3번 던진다고 할 때, 확률변수 X 를 앞면의 수라고 정의하자. 이때 확률변수 X 가 취할 수 있는 값은 0, 1, 2, 3이다.

각각에 해당되는 확률과 기대값을 구하시오.

- 정답)

$$P\{X=0\} = P\{(T,T,T)\} = (\frac{1}{2}) (\frac{1}{2}) (\frac{1}{2}) = 1/8$$

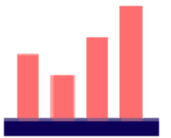
$$P\{X=1\} = P\{(H,T,T), (T,H,T), (T,T,H)\} = 1/8 + 1/8 + 1/8 = 3/8$$

$$P\{X=2\} = P\{(H,H,T), (H,T,H), (T,H,H)\} = 1/8 + 1/8 + 1/8 = 3/8$$

$$P\{X=3\} = P\{(H,H,H)\} = (\frac{1}{2}) (\frac{1}{2}) (\frac{1}{2}) = 1/8$$

$$\text{기대값은 } 0 \cdot P\{X=0\} + 1 \cdot P\{X=1\} + 2 \cdot P\{X=2\} + 3 \cdot P\{X=3\} = 0 + 3/8 + 6/8 + 3/8 = \mathbf{1.5}$$

- 확률이란 어떤 특정한 사건이 일어날 가능성을 0과 1 사이의 값으로 나타낸 것이다.
- 사건 하나하나에 대해서 미리 아는 것은 불가능하더라도 확률적 모형을 통해 많은 시행의 결과를 예측할 수 있다.
- 표본공간과 사건을 벤다이어그램으로 나타내어 특정 사건에 대한 확률을 구할 수 있다.
- 확률변수란 확률 실험으로부터 나타난 결과에 실수를 할당한 함수이다.



Wk4-2 : 빅데이터분석에서의 확률과 분포

- 조건부확률과 베이즈확률 -

1. 조건부 확률과 통계적 독립

- 두 개의 주사위를 던져 두 눈의 합이 10일 확률은 $1/12$ 이다.

그런데 누가 첫번째 주사위의 눈이 4라는 것을 미리 알려주었다!

이때 두 눈의 합이 10일 확률은 몇인가?



1. 조건부 확률과 통계적 독립

4.2 조건부 확률과 베이즈 확률

- 첫 눈이 4라는 것이 주어졌으므로 가능한 결과는

(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) 뿐이다.

이때 두 눈의 합이 10이 되는 사건은 이 중 (4, 6) 이므로 정답은 $1/6$ 이다.

- 왜 결과가 다를까? 표본공간이 달라졌기 때문!

1. 조건부 확률과 통계적 독립

4.2 조건부 확률과 베이즈 확률

- 조건부 확률(conditional probability)

: 어떤 사건(B)이 발생한다는 조건 하에서 다른 사건(A)이 발생하게 될 확률

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- $P(A|B) = P(A)$ 일때, 즉 사건 B가 일어난다는 정보가 사건 A의 발생에

전혀 영향을 주지 않을 때 두 사건이 통계적 독립(independent)이라고 한다.



영국의 철학자
토머스 베이즈 (1701-1761)

확률과 통계를 공부하는 사람이라면 누구나
알고 있는 유명하고, 널리 쓰이고 있는
베이즈 정리 (Bayes' Theorem)!

$$\begin{aligned} P(A_1|B) &= \frac{P(B \cap A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B)} \\ &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} \end{aligned}$$

• 베이즈 정리도 아래와 같은 조건부 확률 계산식으로 볼 수 있다

사건 B가 발생했을 때 사건 A₁가 발생할 확률을
조건부 확률 공식으로 표현

P(B|A₁)에 대한 조건부 확률 공식 이용

$$P(B|A_1) = \frac{P(B \cap A_1)}{P(A_1)}$$

$$P(A_1|B) = \frac{P(B \cap A_1)}{P(B)}$$

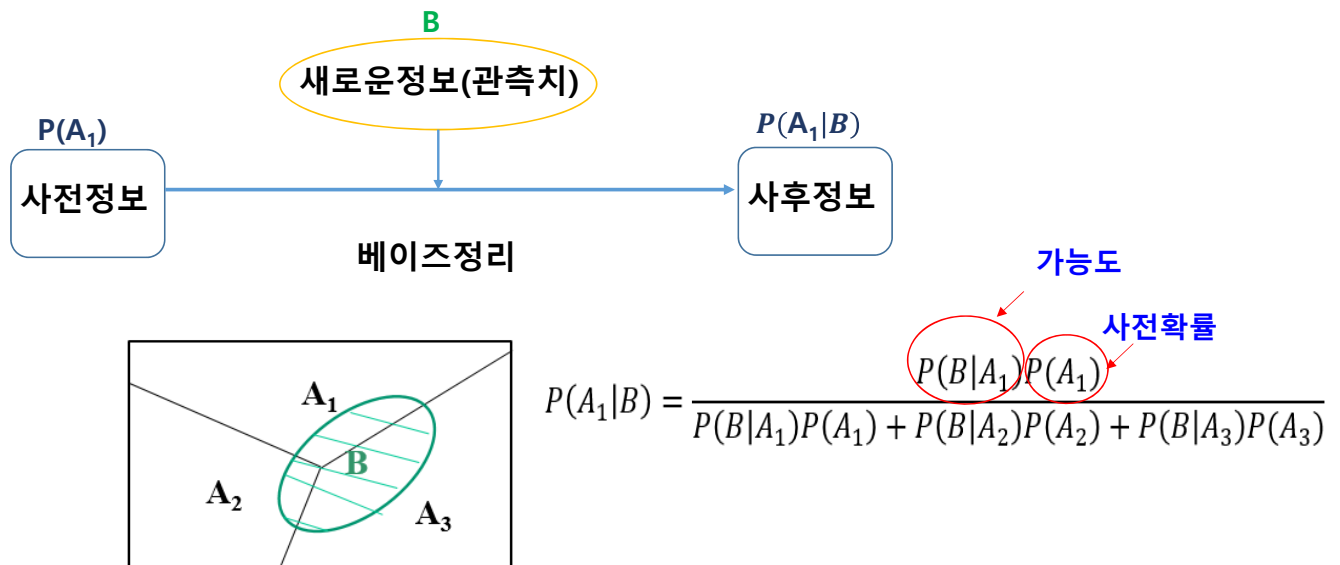
$$= \frac{P(B|A_1)P(A_1)}{P(B)}$$

$$= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$

P(B)를 이와 같이 계산할 수 있다.
P(B) = P(B ∩ A₁) + P(B ∩ A₂)

2. 베이즈 정리

- 주어진(사전정보)가설에 새로운 정보(B)가 주어졌을때 사후확률을 계산한다.



3. 베이즈 확률 - 예제 1

보험회사의 빅데이터 자동 시스템 분석 프로그램에서 사기(fraud)사건으로 의심되면 사기여부에 대한 실제 조사가 실시된다. 실제로 사기사건 확률은 $p(F)=0.15$, 사기사건이 아닌 확률은 $p(F \text{ not})=0.85$ 이다.

그리고 실제 사기사건에 대한 정보(가능도) $p(\text{Pos}|F)=0.95$, $p(\text{Pos}|\text{not } F)=0.01$ 가 경험적으로 주어져 있다.

$p(F)=0.15$, $p(F \text{ not})=0.85$, $p(\text{Pos}|F)=0.95$, $p(\text{Pos}|\text{not } F)=0.01$

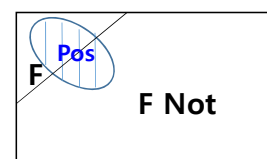
이때 자동 시스템에서 어떤 보험사건이 사기사건이라고 나온 경우 실제 그 사건이 '보험사기'건일 확률은 얼마인지 구해보자.

Fraud (실제 사기사건)

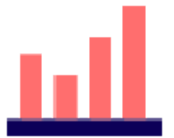
자동 시스템에서 'positive'로 분류

$$\begin{aligned} P(F|\text{Pos}) &= \frac{P(F)P(\text{Pos}|F)}{P(F)P(\text{Pos}|F) + P(\text{not } F)P(\text{Pos}|\text{not } F)} \\ &= \frac{(0.15)(0.95)}{(0.15)(0.95) + (0.85)(0.01)} = 0.9437 \end{aligned}$$

해당 사건은 확률값이 0.9437 이므로 사기사건이라고 분류할수 있다!



- 조건부 확률이란 어떤 사건이 발생한다는 조건 하에서 다른 사건이 발생하게 될 확률이다.
- 베이즈 정리란 사후확률을 사전확률과 가능도를 이용하여 계산할 수 있도록 해 주는 확률 변환식이다.
- 머신러닝기법 중 '나이브베이즈 분류' 기법 계산에서 베이즈정리가 활용된다.



Wk4-3 : 빅데이터분석에서의 확률과 분포

- 정규분포(연속형)와 포아송분포(이산형) -

1. 확률분포란?

- 만약 확률변수 X 가 1, 2.5, 4의 값을 갖고

각각에 대한 확률이 $P(X=1) = 1/2$, $P(X=2.5) = 1/3$, $P(X=4) = 1/6$ 라면?

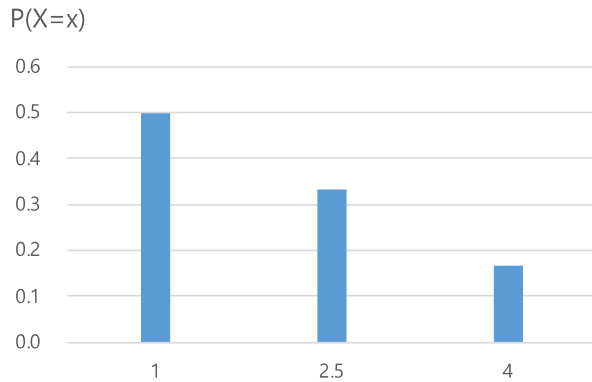
x	1	2.5	4	계
$P(X=x)$	1/2	1/3	1/6	1

이렇게 표로
나타낼 수도 있지만...

1. 확률분포란?

- 만약 확률변수 X 가 1, 2.5, 4의 값을 갖고

각각에 대한 확률이 $P(X=1) = 1/2$, $P(X=2.5) = 1/3$, $P(X=4) = 1/6$ 라면?



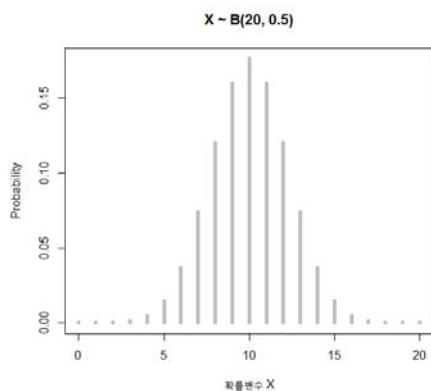
이렇게 함수와 그래프로
나타낼 수도 있다!



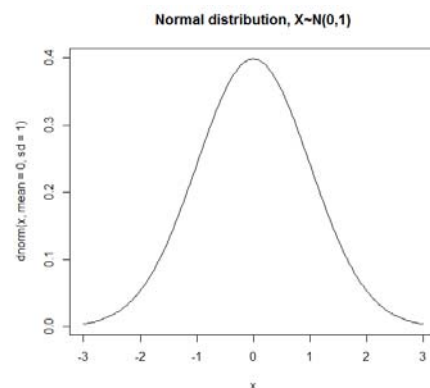
확률변수에 대한 분포 = 확률분포

1. 확률분포란?

- 확률분포에는 이산형(discrete)분포와 연속형(continuous)분포가 있다.



점이 **띄엄띄엄** 분포 되어있음!
→ **이산형 분포**라고 함!



점이 **연속적으로** 분포 되어있음!
→ **연속형 분포**라고 함!

2. 이산형 분포

- 이산형 분포란 **확률변수가 이산형(discrete)**일 때의 확률분포이다!
- 이산확률변수 X 가 특정한 값(x)를 취할 확률 값을 $p(x)$ 로 표기할 때

$$(1) \text{ 기대값 } E(X) = \sum x \cdot p(x) \text{ (가중치 평균의 개념)}$$

$$(2) \text{ 분산 } \text{Var}(X) = \frac{(\bar{x}-x_1)^2 + (\bar{x}-x_2)^2 + \dots + (\bar{x}-x_n)^2}{n-1}$$

$$= \sum_{all\ x} [x - E(X)]^2 \cdot p(x)$$

$$= E(X^2) - E(X)^2$$

2. 이산형 분포 - 예시

(예제 1)

어떤 부품의 직경에 대한 생산공정의 목표치는 10(mm)이다.

다음은 각 직경을 가진 부품이 생산될 확률이다.

직경	9	9.5	10	10.5	11
확률	0.05	0.25	0.4	0.25	0.05

이때 한 부품 직경의 **기대값과 분산**을 구해보자!

(답)

$$E(X) = \sum x \cdot p(x) = (9 * 0.05) + (9.5 * 0.25) + \dots + (11 * 0.05) = 10$$

$$E(X^2) = \sum x^2 \cdot p(x) = (9^2 * 0.05) + (9.5^2 * 0.25) + \dots + (11^2 * 0.05) = 100.225$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = 100.225 - 100 = 0.225$$

3. 이산형 분포 - 이항 분포

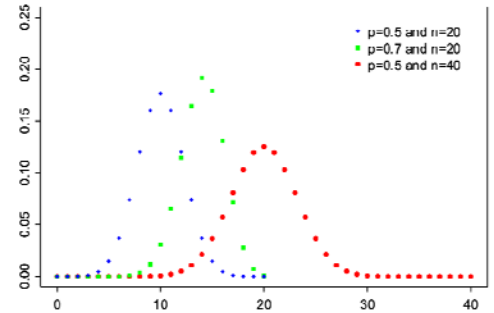
4.3 정규분포(연속형)와 포아송분포(이산형)

- 이산형 분포에는 **이항분포**, 다항분포, 초기하분포, 포아송분포 등이 있다.
- 어떤 시행의 결과가 단순히 '성공' 또는 '실패'로 나타날 수 있을 때 (베르누이시행), 이항분포란 '성공'이 나오는 횟수에 대한 확률분포이다!
- 성공확률이 p 인 베르누이시행을 n 회 반복할 때 성공의 횟수 X 에 대하여

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

$$E(X) = np$$

$$Var(X) = np(1-p)$$



4. 이산형 분포 - 포아송 분포

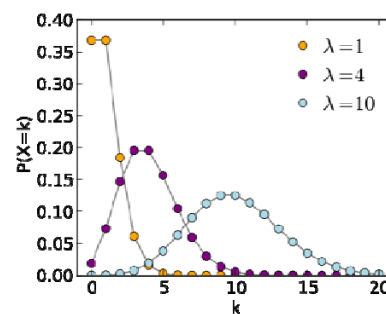
4.3 정규분포(연속형)와 포아송분포(이산형)

- 포아송 분포란 단위 시간 안에 어떤 사건이 몇 번 발생하는가에 대한 확률분포이다!
- 확률변수 X 가 포아송확률변수이고, 모수(평균발생횟수)가 λ 라면

$$f(x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}$$

$$E(X) = \lambda$$

$$Var(X) = \lambda$$



λ 값에 따른 포아송 확률분포

4. 이산형 분포 - 포아송 분포

포아송 분포는 다양한
현실세계의 문제와 밀접한 관련이 있다!

- 일정 주어진 시간 동안에 도착한 고객의 수
- 일정 주어진 생산시간 동안 발생하는 불량 수
- 1킬로미터 도로에 있는 흙집의 수
- 어떤 특정 량의 방사선을 DNA에 쬔었을 때 발생하는 돌연변이의 수

어떤 사람은 하루에 전화를 평균 10회 받는다.
어느 날 이 사람이 6번의 전화를 받을 확률은?

하루 평균 10회의 전화를 받으므로 $\lambda = 10$.
전화 받는 횟수를 확률변수 X 라고 하면,

$$P(X = 6) = f(x) = \frac{e^{-10} \cdot 10^6}{6!}$$

$$= 0.063$$

5. 연속형 분포란?

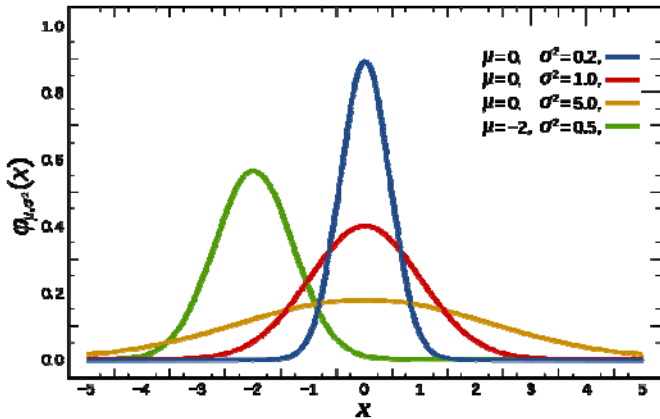
- 연속형 분포란 확률변수가 연속형(continuous)일 때의 확률분포이다.

$$(1) \text{ 기대값 } E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (2) \text{ 분산 } \text{Var}(X) = E(X^2) - E(X)^2$$

- 연속형 분포에서는 정규분포(Normal distribution)가 가장 중요하다!
 - 모집단의 분포가 정규분포를 가진다고 가정하면 여러가지 통계분석이 쉬워짐
 - 실제로 사회적, 자연적 현상의 통계치들의 분포가 정규분포와 비슷한 형태를 띠

5. 연속형 분포 - 정규분포

4.3 정규분포(연속형)와 포아송분포(이산형)



다양한 평균과 분산의 정규분포

- 정규분포는 평균을 중심으로 대칭을 이루는 종모양의 연속확률분포이다.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$X \sim N(\mu, \sigma^2)$$

5. 연속형 분포 - 표준정규분포

4.3 정규분포(연속형)와 포아송분포(이산형)

- 표준정규분포는 평균이 0이고 분산이 1인 정규분포이다.
- 정규분포를 표준정규분포로 만드는 방법은 간단하다!

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- 왜 표준화를 하는가?

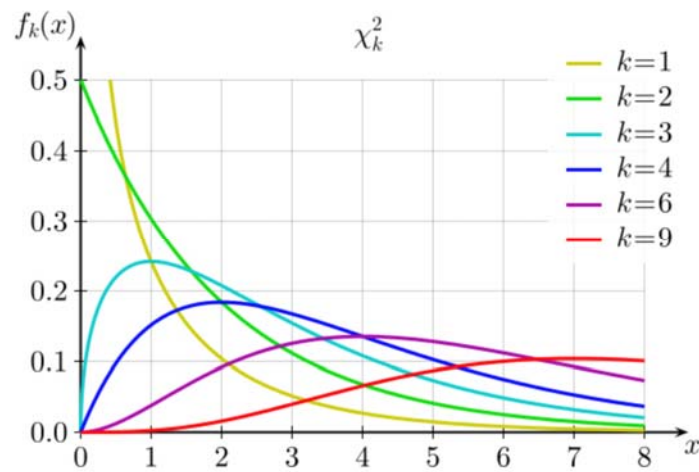
표준정규분포에서의 구간의 면적을 미리 구해두면

이것을 이용해서 모든 정규분포의 면적을 바로 구할 수 있다!

5. 연속형 분포 - 카이제곱(χ^2)분포

4.3 정규분포(연속형)와 포아송분포(이산형)

- 확률변수 Z 가 표준정규분포 $N(0,1)$ 을 따를 때 Z^2 은 자유도가 1인 카이제곱분포를 따른다.



자유도에 따라 달라지는 카이제곱분포의 모습

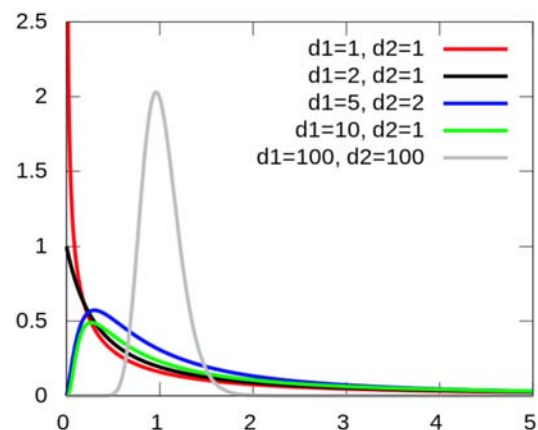
5. 연속형 분포 - F-분포

4.3 정규분포(연속형)와 포아송분포(이산형)

- 두 확률변수 χ_1^2 와 χ_2^2 이 서로 독립이며, 각각의 자유도가 v_1, v_2 인 카이제곱분포를 따를 때, 다음과 같이 정의되는 확률변수 F 는 자유도가 (v_1, v_2) 인 F-분포를 따른다.

$$F = \frac{\chi_1^2/v_1}{\chi_2^2/v_2}$$

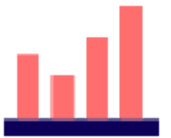
$$F \sim F(v_1, v_2)$$



자유도에 따른 F-분포

6. 요약

- 이산형 분포란 확률변수가 이산형일 때의 확률분포이다.
- 이항분포란 베르누이시행에서 '성공'이 나오는 횟수에 대한 확률분포이다.
- 포아송 분포란 단위 시간 안에 어떤 사건이 몇 번 발생하는가에 대한 확률분포이다.
- 연속형 분포란 확률변수가 연속형일 때의 확률분포이다.
- 정규분포는 정규분포는 평균을 중심으로 대칭을 이루는 종모양의 연속확률분포이다.



Wk4-4 : 빅데이터분석에서의 확률과 분포

- 데이터에서 출발하는 확률과 분포 -

1. 데이터의 분포를 아는 경우

학생 번호	아침식사 여부	흡연여부	동아리활동 시간	범주
1	0	0	2	보통
2	1	0	5	우수
3	0	1	0	보통
4	1	0	4	우수

⋮

- 대학생 20명을 대상으로 생활습관에 대한 설문조사를 하였다.
- 그리고 이 학생들의 특성을 기준으로 **‘보통(0)’**과 **‘우수(1)’** 학생으로 구분하였다.
- 한 학생의 생활습관이 주어졌을 때 이 학생을 어떻게 **분류**할 수 있을까?

1. 데이터의 분포를 아는 경우

- 학생들은 타겟값이 0 또는 1인 **이항분포**를 따름
- 학습 표본을 기반으로 **분류** 규칙을 생성

→ **로지스틱 회귀분석!**

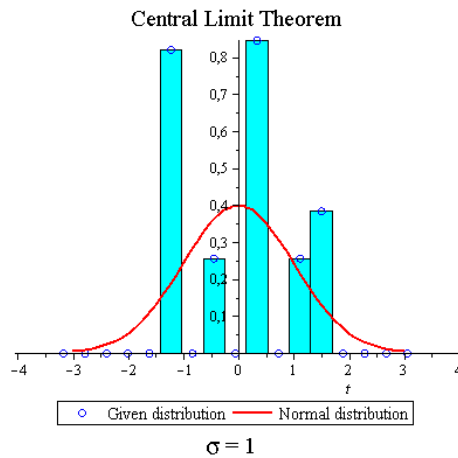
- **데이터의 분포(특성)**를 알고 있다면,
목적에 따라 데이터를 분석하는 것이 수월해진다!

1. 데이터의 분포를 아는 경우

- 하지만 현실세계의 데이터 분포는 무수히 많을 정도로 다양하고
우리가 아는 분포(ex. 포아송분포)로 **설명할 수 없는 분포**도 분명히 존재한다.
- 만일 그 많은 다양한 분포들을 **한 종류의 분포로 근사하여 설명할 수 있다면**
어떨까? → **중심극한정리(Central Limit Theorem)!**

2. 중심극한정리 (Central Limit Theorem)

4.4 데이터에서 출발하는 확률과 분포



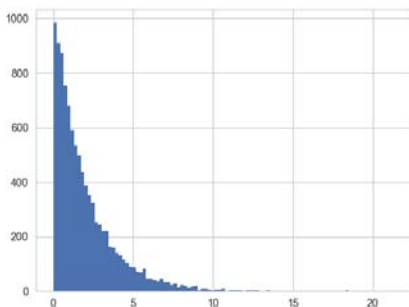
이항분포에서
표본의 수가 증가함에 따라
표본들의 전체 합이
점점 정규분포에 근접해짐!

표본 수의 증가에 따른 전체 합의 확률분포

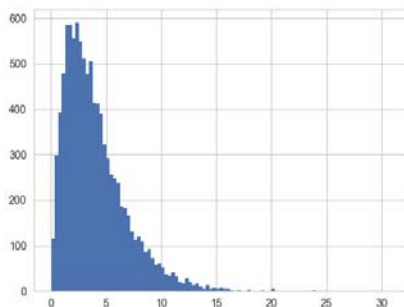
2. 중심극한정리 (Central Limit Theorem)

4.4 데이터에서 출발하는 확률과 분포

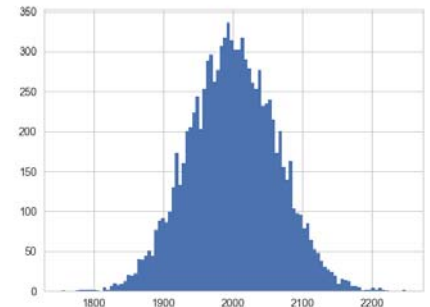
- 지수분포에서도 표본 수의 증가에 따른 표본평균의 분포는
점점 정규분포와 비슷해짐을 볼 수 있음!



X를 10000번 뽑았을 때 ($\lambda=2$)



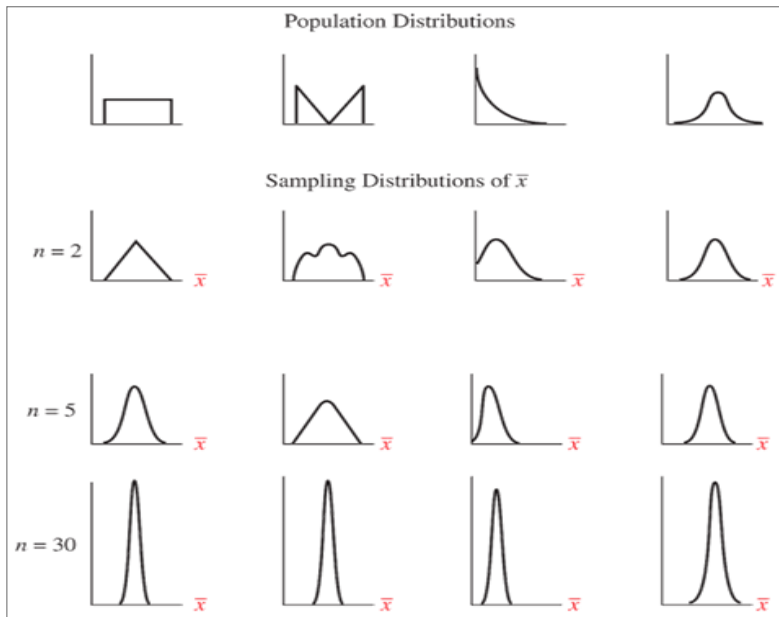
X1, X2를 각각 10000번 뽑고
평균을 구했을 때



X1, ..., X1000를
각각 10000번 뽑고
평균을 구했을 때

2. 중심극한정리 (Central Limit Theorem)

4.4 데이터에서 출발하는 확률과 분포



Agresti, A., Franklin, C., Statistics-The Art and Science of Learning from Data, 3rd ed

원래의 분포가 정규분포가 아니더라도
표본의 수가 증가함에 따라
표본평균이 점점
정규분포모형과 비슷해짐!

2. 중심극한정리 (Central Limit Theorem)

4.4 데이터에서 출발하는 확률과 분포

• 중심극한정리

: 모집단이 정규분포가 아닌 경우에도

표본의 수가 증가하면 표본평균의 분포가 정규분포에 근접한다.

- 평균이 μ 이고 분산이 σ^2 인 모집단으로부터 크기 n 인 확률표본을 추출할 때,
 n 이 크면 표본평균 \bar{X} 는 $N(\mu, \sigma^2/n)$ 에 근접한다.
- 보통 n 이 30 이상이면 모집단의 분포에 관계없이 \bar{X} 는 정규분포에 근사한다.

2. 중심극한정리 (Central Limit Theorem)

4.4 데이터에서 출발하는 확률과 분포

• 중심극한정리가 유용한 이유

대부분의 통계적 검정과 추정은
모집단이 정규분포를 따른다는
가정 하에서 이루어짐



모집단의 분포를 몰라도
중심극한정리를 이용하면 표본평균의
통계적 검정과 추정이 가능해짐!

3. 요약

4.4 데이터에서 출발하는 확률과 분포

- 중심극한정리란 모집단의 분포에 관계없이 표본의 수가 증가하면 표본평균의 분포가 정규분포에 근접한다는 이론이다.
- 평균이 μ 이고 분산이 σ^2 인 모집단으로부터 크기 $n(\geq 30)$ 인 확률표본을 추출할 때 표본평균 \bar{X} 는 $N(\mu, \sigma^2/n)$ 에 근접한다.
- 모집단의 분포를 몰라도 중심극한정리를 이용하면 표본평균의 통계적 검정과 추정이 가능해진다.

참고교재 :

David S. Moore , William I. Notz, 심규박, 조태경, 이승수 옮김, 개념과 논쟁으로 배우는 통계학 2018
Agresti,A., Franklin, C., Statistics, The Art and Science of Learning from Data 3rd ed, 2014

