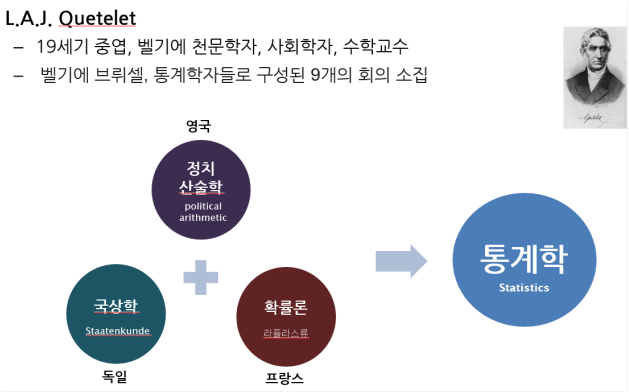
1. 통계학 이해하기

통계학(statistics)

- 데이터 과학의 기초 학문
- 산술적 방법을 기초로 하여 주로 다량의 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 수학의 한 분야(출처. Wikipedia)

역사(유래)

- 라틴어 statisticus(확률), statisticum(상태), 이탈리아어 statista(국가, 나라, 정치가) 등에서 유래
- state + -istic(of, relating to) + -s
- state + -itics(학문)
- 국가의 인력, 재력 등 국가적 자료를 비교 검토하는 학문
- 근대 통계학 : 아돌프 케틀레(Quetelet)에 의해 벨기에의 브뤼셀에서 통계학자들로 구성된 9개 회의 소집을 기원



통계(statistic, 統計)

- 사람, 사물, 사건, 사회적 현상 혹은 자연 현상 등을 조사해 수집된 각종 데이터의 요약
- 집단현상에 대한 구체적인 양적 기술을 반영하는 숫자(출처. 두산백과)

왜 통계학을 알아야 할까?

- 영국 SF소설작가 Hebert George Wells(1866~1946)
  - "읽기 쓰기 능력과 마찬가지로 통계학적 사고 역시 사회인이 갖추어야할 기본 교양이 될 것이다"

[예] A 가전제품 판매 회사

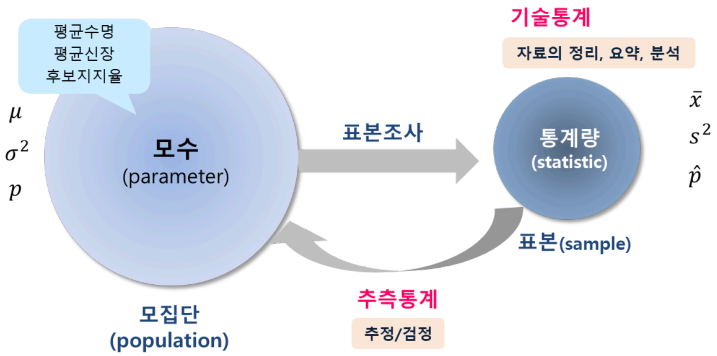
- 매달 DB에 등록된 모든 회원에게 SMS 발송
  - 연 매출 : 10억원
  - SMS 발송비용 7억원
  - 이익 : 3억원
- (의사결정필요) 모든 회원에게 SMS를 보낼 것인가?
- 판촉 효과가 높을 것으로 예상되는 특정 고객에게만 SMS 발송
  - 연 매출 : 8억원

- SMS 발송비용 : 3억원
- 이익 : 5억원

통계학의 목적

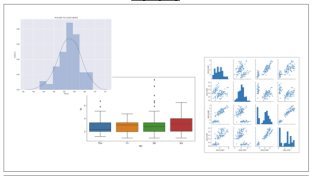
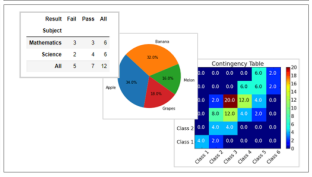
- 학문적 관점
  - 새로운 질문들, 연구과제에 대하여 과학적으로 답을 찾아가는 과정
- 비즈니스 관점
  - 성공 가능성을 높이거나 실패 가능성을 낮추며, 의사결정을 지원
  - 넘치는 데이터로부터 인사이트를 얻는 방법을 제공
  - 빅데이터, 인공지능 시대 기본 학문

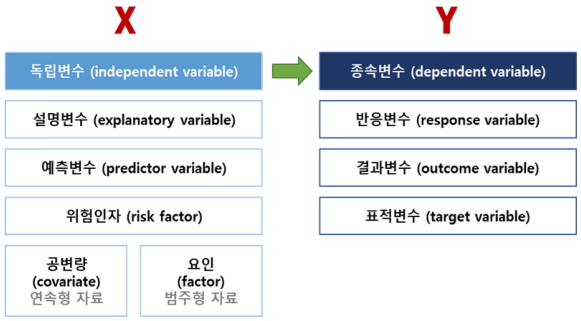
기술통계와 추측통계



기술통계(descriptive statistics)	추측통계(inferential statistics)
• 수집한 데이터를 요약, 묘사, 설명하는 기법 예. 인구조사, 토지조사 등을 통한 현상 파악 • 시각화도구 : 도수분포표, 히스토그램, 상자그림표, 산점도, 버블차트, 히트맵, 평행좌표플롯 등 • 기술통계량 : 평균, 중위수(중앙값), 사분위수,분산, 표준편차, 변동계수, 왜도, 첨도 등	• 수집한 데이터를 기반으로 모집단의 특성을 추론 예측하는 기법 • 전체를 파악할 수 없을 정도의 큰 대상이나 아직 발생하지 않은 미래의 일에 대해 추측하는 기술 예. 대선의 당선 확실 예측, 주가예상, 금융/보험 상품의 가격 결정 등 • **확률이론 기반** • **가설검정 기반**의 통계적 분석 기법 들 - 상관분석, 연관분석, 독립성검정, - 차이검정, 회귀분석, 구조방정식 등

데이터 타입과 역할에 따른 분석 기법

	기술통계	시각화
수치형 (Numerical data)	**"분포 분석"** • 데이터의 특성을 분포로 설명 • 주요 항목은 범위, 평균, 분산, 표준편차 등 • 대부분의 분석 방법이 특정 분포를 가정 • 대표(중심경향), 산포, 왜도, 첨도	
범주형 (Categorical data)	**"빈도 분석"** • 범주별 출현 빈도에 기반한 분석 • 주요 항목은 빈도, 비율, 누적비율 등 • 특정 분포 가정 없이 빈도에 기반한 확률을 사용	

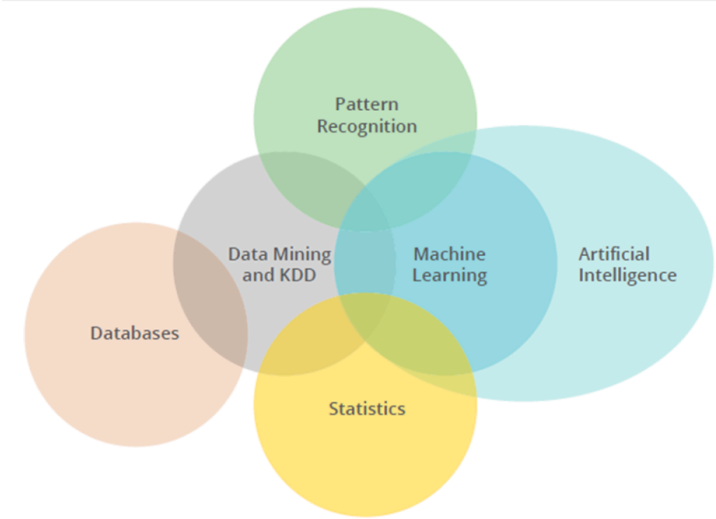


		X (독립변수)	
		수치형	범주형
Y (종속변수)	수치형	상관분석 회귀분석	t-test ANOVA
	범주형	로지스틱 회귀분석	카이제곱검정

데이터 과학과 통계학

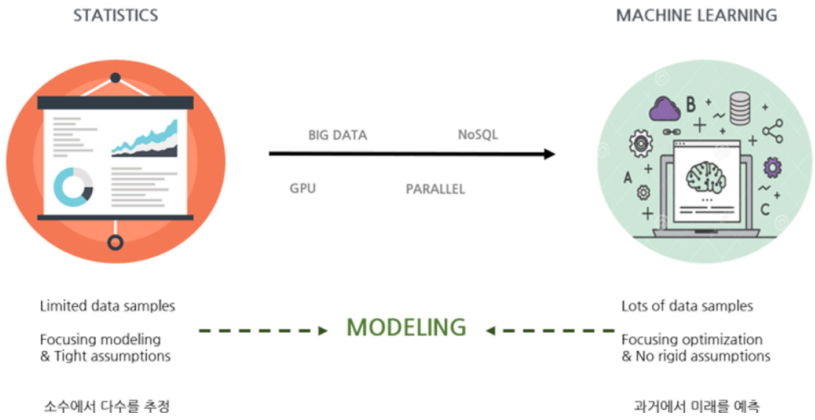
- 데이터 과학은 통계이론으로부터 발전

<데이터 과학 프로세스>



출처 : https://blogs.sas.com/content/subconsciousmusings/files/2014/08/data-mining-Venn-diagram-300x184.png

구분	스몰데이터 분석	빅데이터 분석
분석 목적	• 표본분석을 토대로 모집단의 특성을 추론하여 해석함 • 비교집단간 특성차이,조치수단간 효과차이 비교를 통해 관심대상 집단·수단을 선별하는데 중점	• 대규모 데이터에 숨어있는 패턴을 발견하고 규칙을 도출함 • 도출된 패턴 및 규칙을 이용해 개별대상별 (고객·제품 등) 액션방안 마련에 중점
데이터 수집	• 사용자를 대상으로 한 서베이나 포커스그룹 인터뷰 활용으로 상당한 조사기간이 수반됨	• 대규모 내부거래처리 데이터, 음성, 동영상, 외부공공 및 소셜데이터 활용으로 실시간에 가까운 분석
분석 기법	• 기술통계 및 차이분석, 회귀분석, 구조방정식 등 추론통계 중심	• 군집분석, 연관분석, 분류분석, 예측분석 등 정형 및 텍스트 마이닝 기법이 추가됨
분석 도구	• 전통적 SPSS, SAS 등 상용 통계분석 패키지 활용	• R, 파이썬, 하둡 등의 오픈소스 및 클라우드 기반의 분석도구 활용
적용 분야	• 제조, 금융, 유통, 관광, 보건, 행정, 국방 등 공공 및 민간 전분야를 망라함	• 동일한 적용 분야를 가지며, 보다 다양한 데이터 소스 및 분석 기법 활용으로 분석 근거의 정확성과 예측력이 향상됨
분석 사용자	• 통계학 및 마이닝 전문가에게 의뢰를 통한 분석: Analysis	• 업무실무자 스스로 셀프 분석을 통한 의사결정에 활용: Analytics



출처: https://davincilabs.ai/blog/?q=YToyOntzOjEyOiJrZXl3b3JkX3R5cGUiO3M6MzoiYWxsIjtzOjQ6InBhZ2UiO2k6Mzt9&bmode=view&

---

2. 모집단과 표본추출

모집단과 표본, 전수조사와 표본조사

- 모집단(population)
  - 분석대상 전체 집합(목표모집단)
  - 조사 대상이 되는 관측 가능한 개체로 된 집단 전체(조사모집단)
- 표본(sample)
  - 모집단에서 선택된 모집단 구성단위의 일부

- 전수조사(survey) : 모집단의 자료 전체를 조사
- 표본조사(sampling) : 모집단의 일부를 조사

표본조사를 하는 이유

사례. 1936년 미국 대선

- 공화당 Alfred Landon vs. 민주당 Franklin Roosevelt

구분	Literary Digest	George Horace Gallup
	1916년부터 4차례 미국대선 결과 정확히 예측 (1924년 1600만명, 1928년 1800만명 조사)	대학교수/광고회사 임원 미국여론연구소 설립
조사방법	1000만명 설문지 우송(4.5명에 1명꼴) 237만명 응답(집계만 3개월 소요) 잡지구독자, 전화번호부, 자동차소유자, 대학동창회 명부 이용	비례층화 표본추출법 1500명 조사
조사결과	랜던 후보 57% 예측	루즈벨트 후보 56% 예측
실제투표결과	루즈벨트 60.8%, 랜던 36.	5

- 1952년 이후 8144명 이상의 표본을 사용하지 않음

표본조사 이유

- 표본 데이터를 활용하는 것이 경제적, 시간적으로 유리
- 무한 모집단인 경우

적절한 표본의 수는?

- 통계학자들의 시뮬레이션 결과
  - 일반적으로 최소 200개 이상의 표본이 확보되면 분석이 가능
  - 그러나, 변수 개수나 표본분산에 따라 더 많은 표본이 필요할 수 있음
  - 통계적으로 변수 하나 당 최소 30개의 관측치가 필요
- 데이터 과학을 위한 모델링 단계
  - 수많은 테스트와 검증을 수행
  - 분석모델이 완성될 때까지 표본 데이터 활용하는 것이 경제적, 시간적으로 유리

참고. 표본조사를 통해 모집단의 크기 유추 방법

- 포획-재포획(capture-recapture) 또는 관찰-재관찰(sight-resight)법

오차의 종류

	표본오차(sampling error)	비표본오차(non-sampling error)
정의	모집단과 표본의 자연 발생적인 변동	조사설계나 데이터 수집 및 처리 과정에서 발생되는 오차
원인	표본이 모집단을 완벽하게 대표하지 못하는 무작위성 때문에 발생	잘못된 데이터 수집 방법, 응답자의 응답 오류, 데이터 입력 오류, 표본의 편향(bias)
관리방법	표본의 크기를 증가시키면 표본추출오차는 감소, 표본추출방법 개선하여 무작위성과 대표성을 높임	철저한 조사 설계, 정확한 데이터 수집 및 처리 방법의 선택, 응답자 교육 및 동기 부여 등

표본추출 과정에서 발생하는 편향(bias)

- 표본추출편향(sample selection bias)
  - 표본 추출 과정에서 체계적인 경향이 개입되어 모집단에서 편향된 표본만 추출되는 경우
- 가구편향(household bias)

    - 모집단의 부분 집단 단위에서 하나의 관측치씩 추출하는 경우 크고 적은 집단이 작고 많은 집단보다 적게 추출되는 경우
- 무응답 편향(non-response bias)
    - 설문에 응답하지 않는 사람들과 응답하는 사람들에 체계적이 차이가 있는 경우
- 응답 편향(response bias)
    - 설문 형식의 문제, 응답자의 심리적 이슈에 의해 표본이 영향을 받는 경우
    - 예. 선거당일 출구조사에서 사회적 시선, 여론의 분위기 때문에 거짓을 말하여 편향이 발생(브래들리 효과)

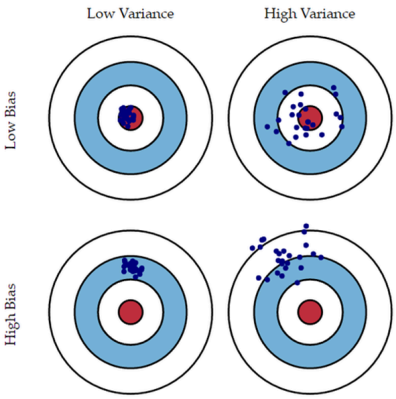
인지적 편향(Cognitive bias)

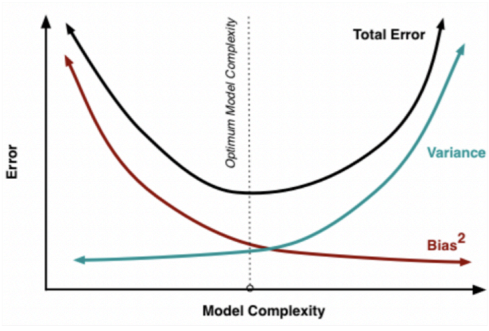
- 분석가의 성향이나 상황에 따라 논리적인 추론을 내리는 패턴

인지적 편향의 종류

- 확증 편향(confirmation bias)
    - 자신이 본래 믿고 있는 대로 정보를 선택적으로 받아들이고 임의로 판단하는 편향
    - 자신의 판단에 대한 확신을 더해주는 방향으로 데이터를 조정
    - 두 명 이상의 분석가가 크로스 체크
- 기준점 편향(anchoring bias)
    - 분석가가 처음에 접하는 정보에 지나치게 매몰되는 편향
    - 처음 표본을 통해 나왔던 통계가 머릿속에 각인되어 다른 분석 결과를 무시하거나 과소평가
- 선택 지원 편향(choice-supportive bias)
    - 의사 결정을 내리는 순간 그 선택의 긍정적인 부분에 더 많이 생각하고 그 결정에 반대되는 증거를 무시하게 되는 편향
    - 기존의 상식과 고정관념으로 정보와 근거들을 선택적으로 수용
- 분모 편향(denominator bias)
    - 분수 전체가 아닌 분자에만 집중하여 현황을 왜곡하여 판단하는 편향
    - 8세기 중국 '안록산의 난' 사상자 4천만명(15%) vs. 1940~1950년 2차세계대전 사상자 25억명(4%)
    - 1990년 금리 10% vs. 2020년 금리 2.5% (물가상승률 1990년 6%, 2020년 0.5%)
- 생존자 편향(survivorship bias)
    - 소수의 성공한 사례를 일반화된 것을 인식함으로 나타나는 편향
    - 예. 2차 세계대전 전투기 총탄 자국 분석하여 취약한 부분 보강
        - 날개와 꼬리 부분에 총탄 자국이 많음
        - 조종석과 엔진 부분의 총탄 자국이 적으나 치명적

참고. 머신러닝 모델 측면에서 편향과 분산





표본 추출 단계

단계	내용	
1단계	모집단 확정	조사대상(사람, 사물, 조직, 지역 등) 전체 집합을 구체적으로 정의
2단계	표본 프레임 결정	조사대상 목록 설정
3단계	표본 추출 방법 결정	확률표본추출/비확률표본추출, 복원/비복원
4단계	표본 크기 결정	조사 유형, 시간, 예산 등 고려
5단계	표본 추출	조사 대상 추출

확률 표본 추출 방법

추출법	내용
단순임의추출 (Simple Random Sampling:SRS)	무작위 추출 모집단 모든 구성단위가 표본으로 선정될 확률이 동일
층화추출 (Stratified Random Sampling)	모집단이 특정한 기준으로 분류가 가능할 때 몇 개의 층(strata)으로 나눔 층 간에는 차이가 존재하므로 각 층에서 골고루 개체 선택(SRS)
계통추출 (Systematic Sampling)	모든 구성 단위에 일련번호를 부여한 뒤 일정한 간격으로 표본을 선택하는 방법
집락추출 (Cluster Sampling)	모집단을 여러 소집단(cluster)으로 분류한 뒤 임의의 소집단들을 선택하여 분석 소집단(집락) 간 동질적, 소집단 내 이질적

3. 변수(variable)와 척도(scale)

변수(variable)의 종류

원인(X)	결과(y)
독립변수(independent variable)	종속변수(dependent variable)
설명변수(exploratory variable)	반응변수(response variable)
입력변수(input variable)	출력변수(outcome variable)
조작변수(manipulated variable)	측정변수(measures variable)
예측변수(predictor variable)	비예측변수(predicted variable)
특징(feature)	표적변수(target)

변수 관계

변수 관계	내용
독립관계	변수 간에 상관성 즉, 상관계수가 0인 관계
상관관계	변수 간에 관련성이 존재하는 관계, 양(+)/음(-)의 관계
인과관계	독립변수 변화가 종속변수 변화에 영향을 주는 경우
쌍방향적 인	인과성이 쌍방으로 미치는 경우, 원인과 결과가 동시에 될 수 있음

변수 관계	내용
과관계	
조절관계	독립변수와 종속변수 사이에서 강하고 불확정적인 영향을 미치는 관계 독립변수A가 종속변수에 미치는 영향력이 독립변수B(조절변수)에 따라 다른 경우
매개관계	독립변수의 결과이면서 동시에 종속변수의 원인이 되는 매개변수가 개입되어 독립변수의 영향을 종속변수에 전달하는 관계

척도(scale)

- 명목척도(nominal scale)
- 서열척도(ordinal scale)
- 등간척도(interval scale)
- 비율척도(ratio scale)

4. 데이터의 기술 통계

기술통계 : 전체 데이터를 쉽고 직관적으로 파악할 수 있도록 설명

- 중심경향(location)
  - 산술평균
  - 가중평균
  - 기하평균
  - 조화평균
  - 중앙값
  - 최빈값
  - 사분위수
- 산포도(dispersion)
  - 분산, 표준편차
  - 범위
  - 사분위간 범위
  - 변동계수
- 왜도와 첨도(shape)
