Introduction to Data Analytics

https://product.kyobobook.co.kr/detail/S000001766480

데이터 분석의 목적

: 수집한 데이터를 사람이 해석하고 이용할 수 있는 형태로 변환해서 분석 대상을 이해하거나 예측하는 것

관측

: 분석 대상으로부터 데이터를 수집하는 것

편향

: 관측 과정에서 여러 가지 의미로 왜곡

: 선택편향, 표본 편향, 자발적 참여자 편향,

- 묵인 응답 경향, 중심화 경향, 캐리오버 경향, 답을 유도하는 질문
- 유리한 데이터만 수집, 인위적인 실수, 단위 오류, 입력 오류

데이터 분석 결과는 수집한 데이터의 품질이 그대로 나타남

개념적 정의

오차 = 값에 변동이 있다

: 우연 오차, 편향 오차, 계통 오차

확률 변수

: 무작위로 얻어진 값을 가지는 변수

확률 분포

확률 밀도

이산적인 값 / 연속적인 값

정규 분포

매개 변수

매개변수 추정 / 피팅

통계 모델링

중심극한정리

표본 평균

대수의 법칙

= 측정 횟수를 늘리면 표본 평균은 참값에 가까워짐

상관계수

상관관계 / 인과관계

겉보기 인과관계

: 우연히 상관관계 발생, 역인과관계, 선택편향이 생기는 가공 실시, 공통 원인이 되는 요인 존재

변수의 얽힘

인과 효과에서의 개입 == 인과적 추론의 근본 문제

무작위 배정 임상시험

: 평균 처치 효과 - 처치군, 실험군 / 대조군, 제어군

수리모델 > 로지스틱 회귀 (- 더미변수)

수리모델

- 선형회귀모델, 통계 모델, 미분방정식 모델, 시계열 모델, 신경망

성향점수 매칭, 군형화, 정규화

표본(샘플링) 조사

: 표본(샘플), 표본 크기, 표본 개수,

: 모집단, 전수조사(전부조사)

무작위 추출

- 단순 무작위 추출, 계통추출법, 충화 다단계 추출법

표본 오차, 신뢰구간

무작위가 아닌 추출

- 유의추출, 편의추출법

일반성 = 외적타당성

표본추출틀

범위 오차

면접조사, 우편조사, 전화조사

인터넷 조사, 집단에 응답 의뢰, 모집에 의한 표본추출

성공한 사람은 응답한다 | 인터넷 이용자 | 귀찮은 응답

데이터는 손을 타면 탈수록 에러가 생긴다

무시 할 수 없는 이상치

- 이상치에 해당하는 부자, 대공황의 발단, 미증유의 자연재해

데이터 해석까지의 흐름

: 데이터 관측 - 전처리 - 분석 - 결과 해석 및 이용

데이터 관측

: 실험이나 조사 실시, 계측 시스템 도입, 공개 데이터 이용

전처리

: 이상치, 결측치, 노이즈 제거, 포맷 조정, 데이터 표준화 및 가고

분석

: 통계 모델링, 머신러닝, 수리 모델 구축

결과 해석 및 이용

: 수리 모델 성능 평가, 기존 지식과 비교 해석 및 이해, 시스템 구축

처리 코드 통일 및 분석 코드 관리, 소프트 웨어 이용

- 알기 쉬운 변수명, 여러번 반복하는 처리는 하나로 정리, 긴 처리는 가능한 분할, 짧게 쓰는 것이 전부는 아니다

데이터 보관

: 데이터 분석 결과를 학술 논문에 발표하는 경우, 그 데이터는 원칙적으로 10년간 보존하는 것이 바람직

대표적인 값

- 기술통계량, 요약 통계량
- 중앙값, 최빈값, 최댓값, 최솟값, 백분위수

기술 통계량

= 분포 전체의 정보를 대략 종합한 것

스트립플롯, 스웜 플롯, 막대 그래프 - 에러바, 바이올린 플롯,

이항분포, 로그 정규분포, 파레토 분포, 레비 분포, 와이블 분포

시계열 데이터

주기 변동

자기 상관

가설 검정, 귀무 가설

가설 설정 > 검정 방법 선택 > 가설 검정 시행

정규성 / 등분산성

F 검정 / 스튜던트 t 검정

- 효과 크기, 코헨의 d

대응 비교

비대응 비교

대응 표본 t 검정

윌콕슨 부호 순위 검정

다변량 데이터

탐색적 데이터 분석

확증적 데이터 분석

다중성 보장

- 본페로니법, 홈법

분산 분석

- 일원 배치 분산 분석

주효과

다중 비교

튜키법

그래피컬 모델링, 경로 분석, 공분산 구조분석, 구조 방정식 모델링

주성분 분석

계층적 군집화

- 덴드로그램

목적 변수, 종속 변수

설명 변수, 독립 변수

수리 모델의 타당성

- 모델 구축에 사용한 데이터를 설명할 수 있는가 : 적합도, 결정계수
- 미지의 데이터를 설명할 수 있는가
- 논리적 타당성

심층학습

과적합, 오버피팅

일반화

초기치 예민성

선형 / 비선형

HARKing

p-hacking

p-hacking 피하고자 만든 가이드 라인

- 데이터 수집 전 어디까지 데이터를 수집할 것인가를 결정하여 보고한다
- 하나의 조건에 최소한 20개의 관측값을 은다
- 수집한 모든 변수에 대해 보고한다
- 데이터를 수집한 모든 실험 조건을 보고한다
- 만약 관측값을 제거하는 경우는 그것을 제거하지 않은 경우의 분석 결과도 표시한다
- 분석에서 어떤 변수의 영향을 제거하는 조작을 한 경우는 그렇게 하지 않은 경우의 결과도 표시한다

힐의 기준

- 강도, 일관성, 특이성, 시간성, 용량과 반응의 관계, 타당성, 정합성, 실험 유무, 유사성

선후 인과의 오루

도박사의 오류

가용성 편향

확증 편향 > 체리 피킹

문맥의 효과