3과목 데이터분석

데이터마트(DM)

데이터 웨어하우스의 한 분야로 특정목적을 위해 사용 (소규모 데이터웨어하우스)

EDA(탐색적 자료 분석)

데이터의 의미를 찾기위해 통계, 시각화를 통해 파악

EDA의 4가지 주제 저향성의 강조, 잔차계산, 자료변수의 재표현, 그래프를 통한 현시성

(저잔재현)

결측값

존재하지 않는 데이터 null로 표시

1)단순대치법

Complet case 함수로 False데이터에 결측값 제거

2)평균대치법

평균으로대치

3)단순확률대치법

가까운 값으로 변경

4)다중 대치법

여러번 대치 (대치 -> 분석 -> 결합)

이상값

이상값을 항상 제거하는것은 아님

1)ESD 평균으로부터 표준편차의 3배가 넘어가는 데이터는 이상값으로 판단

2)사분위수

통계분석

전수조사와 표본조사

전수조사: 전체를 다 조사, 시간과 비용 많이 소모

표본조사: 일부만 추출하여 모집단을 분석

표본 추출 방법

1)랜덤 추출법: 무작위로 표본추출

2)계통 추출법: 번호 부여하여 일정 간격으로 추출

3)집략 추출법: 여러 군집으로 나눈 뒤 군집을 랜덤추출

군집 내 이질적 특징, 군집 간 동질적 특징

4)층화 추출법: 군집 내 동질적 특징, 군집 간 이질적 특징,

같은 비율로 추출 시 비례층화 추출법

5)복원, 비복원 추출법:

자료의 척도구분

등간척도: 구간사이에 간격이 의미가 있으며 덧셈 뺄셈만 가능(온도, 지수 등)

비율척도: 절대적 기준 0이 존재하고 사칙연산 가능한 자료(무게, 나이 등)

Summary함수 결과의 해석

Mean, median 등 존재: 수치형 변수(회귀분석)

집단의 빈도 수: 범주형 변수(분류분석)

기초확률 이론

1)조건확률 특정사건 B가 발생했을 때 A가 발생할 확률

2)독립사건: A,B가 서로 영향을 주지 않는 사건

3)배반사건: A,B가 서로 동시에 일어나지 않는 사건

확률분포

1. 이산 확률 분포

이산균등분포, 베르노이분포, 이항분포, 기하분포, 다항분포, 포아송분포

(배포항항하)

추정

1. 점추정: 모집단이 특정한 값
2. 구간추정: 모집단이 특정한 구간(95%, 99%를 가장 많이 사용)

가설검정

1. 귀무가설: 일반적으로 생각하는 가설(차이가 없다)

차이가 없다 혹은 동일하다

1. 대립가설: 귀무가설을 기각하는 가설, 증명하고자하는 가설(차이가 있다, 크다/작다)
2. 유의수준: 귀무가설이 참일때 기각하는 1종오류를 범할 확률의 허용 한계 일반적 5%
3. 유의확률: 귀무가설을 지지하는 정도를 나타내는 확률

회귀분석

개념: 독립변수들이 종속변수에 영향을 미치는 파악하는 분석방법

1. 독립변수: 원인을 나타내는 변수
2. 종속변수: 결과를 나타내는 변수
3. 잔차: 계산값과 예측값의 차이

회귀계수 추정방법

최소제곱법: 잔차의 제곱합이 최소가 되는 회귀계수와 절편을 구하는 방법