확증적 데이터 분석 (Confirmatory Data Analysis)

* 정의: 미리 설정한 가설을 데이터를 통해 검증하는 분석 접근법
* 예: "COVID-19 백신이 감염 예방에 효과적이다"라는 가설을 임상시험 데이터로 검증

1. 탐색적 데이터 분석 (Exploratory Data Analysis)

* 정의: 데이터의 숨겨진 패턴, 구조, 관계를 발견하기 위한 분석 방법
* 예: 고객 구매 데이터에서 예상치 못한 구매 패턴 발견

탐색적 데이터 분석의 4가지 주제:

2-1. 저항성의 강조

* 정의: 극단값이나 이상치에 덜 민감한 통계적 방법 사용
* 예: 평균 대신 중앙값 사용

2-2. 잔차 계산

* 정의: 모델 예측값과 실제값 사이의 차이를 분석
* 예: 선형 회귀 모델에서 예측 오차 계산

2-3. 자료변수의 재표현

* 정의: 데이터를 다양한 방식으로 변환하여 분석
* 예: 로그 변환, 제곱근 변환

2-4. 그래프를 통한 현시성

* 정의: 시각화를 통해 데이터의 특성과 패턴을 이해
* 예: 상자 그림(Box Plot), 산점도(Scatter Plot)

1. 상관관계 분석

3-1. 피어슨 상관계수

* 정의: 두 연속형 변수 간의 선형적 관계 강도를 측정
* 범위: -1 ~ +1
* 예: 키와 체중 사이의 상관관계

3-2. 스피어만 상관계수

* 정의: 두 변수의 순위 간 상관관계를 측정
* 비선형 관계도 파악 가능
* 예: 교육 수준과 소득 수준의 관계

1. 공분산

* 정의: 두 변수가 함께 변화하는 정도를 측정
* 양수: 같은 방향으로 변화
* 음수: 반대 방향으로 변화
* 예: 광고 지출과 매출 간의 공분산

1. 기초 통계량

5-1. 중앙값

* 정의: 데이터를 정렬했을 때 중간에 위치한 값
* 예: [1, 2, 3, 4, 5]의 중앙값은 3

5-2. 최빈치

* 정의: 가장 자주 등장하는 값
* 예: [1, 2, 2, 3, 4]의 최빈치는 2

5-3. 사분위수

* 정의: 데이터를 4등분하는 지점
* 예: 25% 지점(1사분위), 50%(중앙값), 75%(3사분위)

1. 평균과 중앙값 비교

* 정의: 데이터의 분포 특성 이해
* 예: 극단값에 따라 평균과 중앙값이 다를 수 있음

1. 산포도

* 정의: 데이터가 얼마나 퍼져 있는지 측정
* 예: 분산, 표준편차

1. 시공간 데이터 탐색

* 정의: 시간과 공간 차원을 포함하는 데이터 분석
* 예: 지역별 인구 변화, 기후 변화 추적

1. 다변량 데이터 탐색

* 정의: 여러 변수를 동시에 분석
* 기법: 주성분 분석, 선형판별분석 등
* 예: 고객 세분화, 주식 포트폴리오 분석

1. 비정형 데이터 탐색

* 정의: 구조화되지 않은 데이터 분석
* 기법: 텍스트 분석, 웹 마이닝
* 예: 소셜 미디어 감성 분석, 고객 리뷰 분석

표본추출

11-1. 모집단

* 정의: 연구 대상이 되는 전체 집단
* 종류:
  + 전수조사: 모든 개체를 조사
  + 표본조사: 일부 개체만 조사
* 예: 국가 전체 인구 중 일부를 선택하여 조사

11-2. 표본추출 방법

* 확률적 표본추출: 무작위로 표본 선택
* 계통추출: 일정 간격으로 표본 선택
* 층화추출: 집단을 층으로 나누어 추출
* 군집추출: 집단을 군집으로 나누어 추출
* 예: 대학생 만족도 조사에서 학과별로 비례 할당

1. 자료측정 방법

* 정의: 데이터를 수집하고 분류하는 방법
* 척도 종류: 명목척도, 서열척도, 등간척도, 비율척도
* 예: 성별(명목), 학년(서열), 온도(등간), 나이(비율)

1. 확률분포

13-1. 이산형 확률변수

* 정의: countable한 값만 가지는 변수
* 예: 주사위 던지기, 동전 앞면/뒷면

13-2. 연속형 확률변수

* 정의: 연속적인 값을 가지는 변수
* 예: 키, 체중

주요 확률분포:

* 베르누이 분포: 성공/실패 이벤트
* 이항분포: 고정된 횟수의 독립적 시행
* 포아송 분포: 희귀한 사건 발생 빈도
* 정규분포: 대칭적 종 모양 분포
* t분포: 작은 표본 크기에서 평균 추정
* 카이제곱 분포: 분산 추정

1. 표본분포

* 정의: 모집단에서 반복 추출한 표본들의 통계량 분포
* 구성 요소: 모집단, 표본, 모수, 통계량
* 예: 여러 번 추출한 표본들의 평균 분포

1. 추론통계

15-1. 통계적 추정

* 점추정: 단일 값으로 모수 추정
* 구간추정: 모수가 존재할 가능성이 높은 범위 추정
* 예: 국민 평균 소득 추정

1. 바람직한 통계적 추정량 기준

* 불편성: 기대값이 참값과 일치
* 효율성: 가장 작은 분산
* 일치성: 표본 크기 증가에 따라 정확성 향상
* 충족성: 모든 유용한 정보 활용

1. 가설검정

* 귀무가설: 기존 가정
* 대립가설: 새로운 주장
* 검정 오류:
  + 제1종 오류: 옳은 귀무가설 기각
  + 제2종 오류: 잘못된 귀무가설 채택
* 예: 신약 효과 검증

1. 분석모형 설계

* 데이터 유형 및 속성 파악
* 연속형/범주형 변수 분석
* 분석 기법:
  + 통계분석: 회귀, 분산분석
  + 데이터 마이닝: 트리 기반, 최적화
  + 머신러닝: 지도/비지도/강화학습

1. 분석절차

* 요건 정의
* 모델링
* 검증 및 테스트
* 모델 적용
* 고려사항:
  + 훈련/테스트 세트
  + 과대/과소적합 방지
  + 하이퍼 파라미터 조정

1. 분석 도구

* R: 통계 분석에 특화
* Python: 머신러닝, 데이터 과학에 강점