기상기후빅데이터

Weather and Climate Big Data

빅데이터 개론 2

오승민

Spring 2025

빅데이터 개론 2

- **빅데이터 탐색**
  - 데이터 전처리
    - 데이터 정제
    - 분석 변수 처리
  - 데이터 탐색
    - 데이터 탐색 기초
    - 고급 데이터 탐색
  - 통계기법 이해
    - 기술통계
    - 추론통계

데이터 전처리

데이터 정제 > 데이터 이해

- **데이터 종류**
  - 단변량 자료 (특성변수가 하나), 다변량 자료 (두 가지 이상)
  - 질적자료 (정성적/범주형), 수치자료, 횡적자료, 종적자료

- **데이터 정제**
  - 수집 데이터에서 분석에 필요한 데이터를 추출하고 통합하는 과정
  - 집계(aggregation): 데이터 요약 (합계, 평균, 분산 등)
  - 일반화(generalisation): 일반적 특성이나 패턴 추출
  - 정규화(normalisation): 데이터를 일정한 범위로 조정, 표준화
  - 평활화(smoothing): 변동 줄이고 노이즈 제거, 추세나 패턴 부드럽게

데이터 전처리

데이터 정제 > 결측값 처리

- **결측치(Missing Data)는 데이터가 없음을 의미**
  - 결측치를 임의로 제거 시, 분석 데이터의 직접 손실
  - 결측치를 임의로 대체 시: 데이터 편향 발생, 신뢰성 저하 가능

- **결측 데이터 종류**
  - 완전 무작위 결측(MCAR): 어떤 변수의 결측 데이터가 관측된 혹은 관측되지 않은 다른 변수와 아무런 연관이 없는 경우
  - 무작위 결측(MAR): 관측된 다른 변수와 연관되어 있지만 비관측값들과는 연관되지 않음
  - 비 무작위 결측(NMAR): 결측 변수값이 결측여부(이유)와 관련이 있는 경우

데이터 전처리

데이터 정제 > 결측값 처리

- **결측치 처리**
  - ▸ 단순 대치법
  - 완전 분석: 불완전 자료는 무시
  - 평균 대치법: 관측/실험한 데이터의 평균으로 결측치 대치
  - 회귀대치법: 회귀분석에 의한 예측치로 대치
  - 단순 확률 대치법: 전체 데이터 중 무작위로 대치
  - 최근접 대치법: 가장 가까운 값으로 대치

  - ▸ 다중 대치법
  - 단순 대치법을 복수로 시행: 대치 -> 분석 -> 결합(평균값 등)

데이터 전처리

데이터 정제 > 이상값 처리

- **이상치/이사값(outlier)은 정상의 범주(전체적 패턴)을 벗어난 값을 의미**
  - 단변수 이상치: 하나의 데이터 분포에서 발생하는 이상치
  - 다변수 이상치: 복수의 연결된 데이터 분포 공간에서 발생하는 이상치

- **이상치 발생원인**
  - 비자연적: 입력실수(수집과정 에러), 측정오류, 실험오류, 의도적 이상치(예: 키 조사), 자료처리오류(전처리 과정 에러), 표본오류(표본 추출 시 편향 발생)
  - 이 외에는 자연적 이상치(실제 존재하는 극단적인 값)

데이터 전처리

데이터 정제 > 이상값 처리

- **이상치 탐지**
  - 시각화: 박스 플랏, 줄기-잎 그림, 산점도
  - Z-score: 통계적 지표
  - 밀도기반 클러스터링: 군집간의 밀도 (비모수적)
  - 고립 의사나무 방법: 랜덤한 분할을 통해 고립된 데이터

데이터 전처리

분석 변수 처리 > 변수 선택

- **변수선택:** 분석 목적에 맞게 유의미한 정보 제공하는 적절한 변수를 선택
  - 변수 별 모형: 전체모형(Full Model), 축소 모형(Reduced Model), 영 모형(Null Model: 독립변수가 하나도 없음)
  - 변수 선택 방법: 전진 선택법, 후진 선택법(소거법), 단계적 선택법

데이터 전처리

분석 변수 처리 > 변수 선택

- **변수선택**
  - 전진 선택법 (Forward Selection): 영 모형에서 시작, 모든 독립변수 중 종속변수와 단순 상관관계 절대값이 가장 큰 변수를 분석 모형에 포함시키는 것. 한번 추가된 변수는 제거하지 않는 것이 원칙
  - 후진 선택법 (Backward Selection): 전체 모형에서 시작, 단순상관관계가 작은 독립변수부터 제외 시킴. 한번 제거한 변수는 다시 추가하지 않는다.
  - 단계적 선택법 (Stepwise): 전진 선택법과 후진 선택법의 보완방법. 전진 선택법을 통해 가장 유의한 변수를 포함 후, 나머지 변수들에 대해 후진 선택법을 적용하여 제거 (즉, 성능 감소하는 경우는 제거).

데이터 전처리

분석 변수 처리 > 차원 축소

- **차원 축소: 분석하는 데이터 변수의 양을 줄이는 것**
  - 복잡도 축소: 분석 시간, 저장 공간 효율성 증대
  - 모델의 과적합 방지: 분석모형의 안정성 (robustness) 증대
  - 해석력 확보: 간단한 모델은 해석이 쉬움
  - 차원의 저주: 학습데이터의 수가 차원의 수보다 적어지면 모델 성능 저하되는 것을 방지

데이터 전처리

분석 변수 처리 > 차원 축소

- **차원 축소 방법**
  - 요인 분석(Factor Analysis): 변수들 간의 관계(상관관계)를 분석하여 공통차원을 축약하는 통계분석 과정, 정보손실을 억제하면서 소수의 요인(Factor)으로 축약
  - 주성분 분석 (PCA: Principal Component Analysis): 분산(Variance)을 최대한 보존하는 방향으로 선형 변환하여 차원 축소
  - 특이값 분해 (SVD: Singular Value Decomposition): 행렬을 저차원 성분(특이값과 직교 행렬)으로 분해하여 중요한 정보만 유지
  - 음수 미포함 행렬분해 (NMF: Non-negative Matric Factorisation): 특징 행렬과 가중치 행렬로 분해하여 차원 축소 (이미지 분해)

데이터 전처리

분석 변수 처리 > 파생변수

- **파생변수와 요약변수**
  - 주어진 원데이터를 그대로 활용하기 보다는 분석의 목표에 적합하게 데이터 형태를 수정 보완
  - 주로 데이터 마트(Data Mart: 데이터 웨어하우스로부터 복제 또는 자체 수집된 데이터 모임의 중간층)에 파생변수와 요약변수 저장
  - 요약변수: 수집된 정보를 분석에 맞게 종합한 변수(단순 종합 개념) 예) 매장이용 횟수, 기간별 구매금액 및 구매 횟수 <- 많은 분석 모델에서 공통으로 사용할 수 있어 재활용성이 높음
  - 파생변수: 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여. 예) 주 구매매장 변수, 주 활동지역 변수 <- 논리적 타당성 필요

데이터 전처리

분석 변수 처리 > 변수 변환

- **변수 변환(transformation)**
  - 데이터를 분석하기 좋은 형태로 바꾸는 작업
  - 수학적 의미로는, 기존의 변수 공간에서는 해결하거나 관찰할 수 없는 사실을 영역을 달리(변환)하여 해석이 용이해지거나 취급이 단순해지는 장점
  - 전처리 과정의 하나로 간주
  - 변수 변환 방법: 범주형 변환, 정규화, 로그변환, 역수변환 등

데이터 전처리

분석 변수 처리 > 변수 변환

- **변수 변환(transformation)**
  - 범주형 변환: 연속형/수치형 변수 -> 범주형 변수
  - 정규화: 데이터 스케일이 차이나는 경우, 같은 범위로 변환
  - 로그변환: 로그를 취함
  - 역수변환: 역수를 사용
  - 지수변환: 지수를 사용
  - 제곱근 변환: 제곱근 사용, 정규 분포에 가깝게 변환
  - Box-Cox: 다양한 지수를 적용한 비선형 변환 (최적방법 찾기)
  - 차원 축소

데이터 전처리

분석 변수 처리 > 불균형 데이터

- **불균형 데이터 처리**
  - 가중치 균형방법: 손실 함수에서 소수 클래스(불균형한 클래스)에 더 큰 가중치를 부여하여 모델이 균형 잡힌 학습을 하도록 유도
  - 언더샘플링: 대표(다수) 클래스의 데이터를 줄여(제거하여) 균형을 맞추는 방법
  - 오버샘플링: 소수 클래스의 데이터를 인위적으로 증가(복제 또는 생성)시켜 데이터 균형을 맞추는 방법

데이터 전처리

분석 변수 처리 > 인코딩

- **인코딩 (Encoding)**
  - 범주형 데이터를 수치형으로 변환하여 머신러닝 모델 등에 적용 가능하게 하는 기법
  - 레이블 인코딩 (Label): 각 범주에 순차적인 정수 레이블 할당, 순서나 크기에 의미가 없는 경우 (예: 옷 사이즈 등)
  - 원-핫 인코딩 (One-Hot): 각 범주에 해당하는 인덱스만 1이고 나머지는 0인 이진 벡터로 변환 (예: 사과: [1,0,0], 바나나:[0,1,0])
  - 타깃 인코딩(Target): 각 범주에 속한 종속 변수(타깃)의 평균값을 인코딩으로 사용, 주로 분류 문제에 사용 (예: 카테고리 A의 판매량 100, 150 -> 125를 타겟으로)

빅데이터 개론 2

- **빅데이터 탐색**
  - 데이터 전처리
    - 데이터 정제
    - 분석 변수 처리
  - 데이터 탐색
    - 데이터 탐색 기초
    - 고급 데이터 탐색
  - 통계기법 이해
    - 기술통계
    - 추론통계

데이터 탐색

데이터 탐색 기초 > 데이터 탐색 개요

- **탐색적 데이터 분석 (EDA: Exploratory Data Analysis)**
  - 수집한 데이터를 관찰하고 이해하는 과정, 본격적인 데이터 분석 전에 직관적인 방법으로 통찰하는 과정
  - 데이터에 내재된 잠재적 문제점을 인식하고 해결안을 도출할 수 있음 (본 분석 전 데이터 수집 결정)
  - 문제정의 단계에서 인지 못한 새로운 양상, 패턴을 발견할 수 (초기 가정 수정)
  - 분석 과정: 분석 목적 및 변수, 개별변수 설명 확인, 데이터 문제성(결측치/이상치) 확인, 관계속성(변수 간 상관관계 등) 확인

데이터 탐색

데이터 탐색 기초 > 상관관계

- **상관관계 분석**
  - 두 변수 간의 관계 정도를 나타내며, 한 변수의 변화가 다른 변수의 변화와 어떻게 연관되는지를 측정하는 통계적 개념
  - 단순상관분석 (simple correlation analysis): 두 변수 간의 상관 관계 분석
  - 다중상관분석 (multiple correlationg analysis): 한 종속변수와 여러 독립변수 간의 관계를 분석
  - 편상관분석 (partial correlation analysis): 두 변수 간 상관관계에서 다른 변수들의 영향을 통제한 후 상관성 분석

데이터 탐색

데이터 탐색 기초 > 상관관계

- **상관관계 분석 방법**
  - 피어슨 상관계수 (Pearson Correlation Coefficient): 두 변수 간의 선형적 관계를 측정하는 상관계수, 값이 1에 가까울수록 양의 상관, -1에 가까울수록 음의 상관, 0에 가까울수록 상관이 없음
  - 스피어만 상관계수 (Spearman Correlation Coefficient): 두 변수 간의 단조(monotonic) 관계를 측정하는 상관계수, 데이터의 순위를 기반으로 계산.

데이터 탐색

데이터 탐색 기초 > 기초통계량

- **기초통계량의 추출**

  - **중심화 경향 기초 통계량 (Central Tendency)**
    - 산술 평균: 전체 합을 전체 자료 수(n)로 나눔, 일반적인 평균
    - 기하평균: 관측치를 곱한 후, n의 제곱근으로 표현 (예: 수익률, 상승률)
    - 조화평균: 관측치의 역수의 산술평균 구한 후, 다시 역수를 취함 (변화율)
    - 중앙값: 크기 순으로 나열할 때 가운데 위치한 값
    - 최빈값: 가장 노출 빈도가 높은 자료
    - 분위수: 자료의 위치를 표현하는 수치 (예: 사분위수)

데이터 탐색

데이터 탐색 기초 > 기초통계량

- **기초통계량의 추출**

  - **퍼짐 정도 (산포도, 분산도)**
  - 분산: 평균을 중심으로 퍼진 정도
  - 표준편차: 분산의 제곱근
  - 범위: 최대값과 최소값의 차이
  - 평균절대편차: 편차(각 자료값과 평균 차이) 절대값의 산술평균
  - 사분위범위: Q3-Q1값, 자료의 50% 범위 내 위치
  - 변동계수(Coefficient of Variance): 평균을 중심으로 한 상대적인 산포의 척도 (평균 다른 두 집단 비교), (표준편차/평균)*100

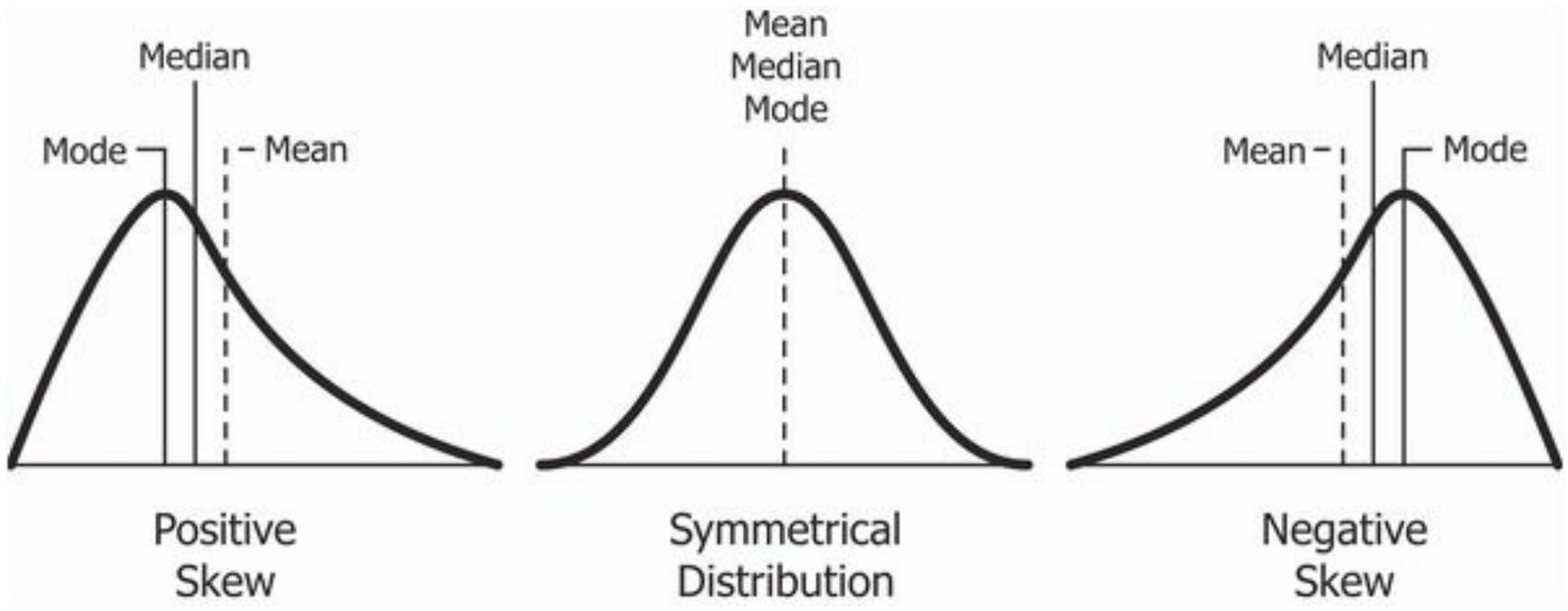
데이터 탐색

데이터 탐색 기초 > 기초통계량

- **기초통계량의 추출**

  - **자료의 분포 형태 (Distribution Shape)**

    - 왜도 (Skewness): 분포의 비대칭 정도를 나타내는 측도

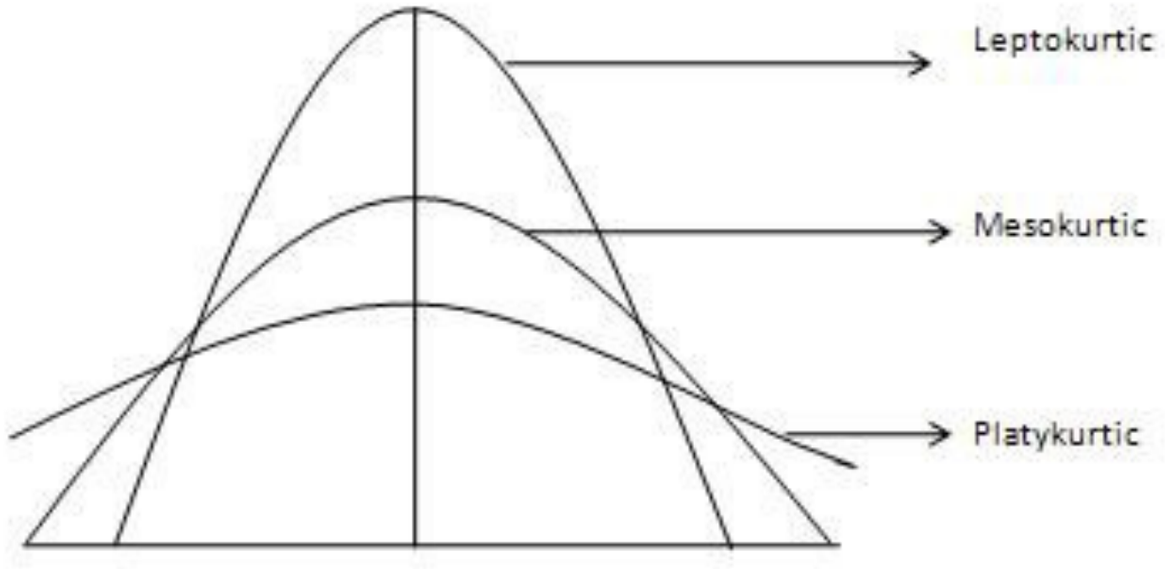


데이터 탐색

데이터 탐색 기초 > 기초통계량

- **기초통계량의 추출**

  - **자료의 분포 형태 (Distribution Shape)**
    - 첨도 (Kurtosis): 분포의 뾰족한 정도를 나타내는 척도

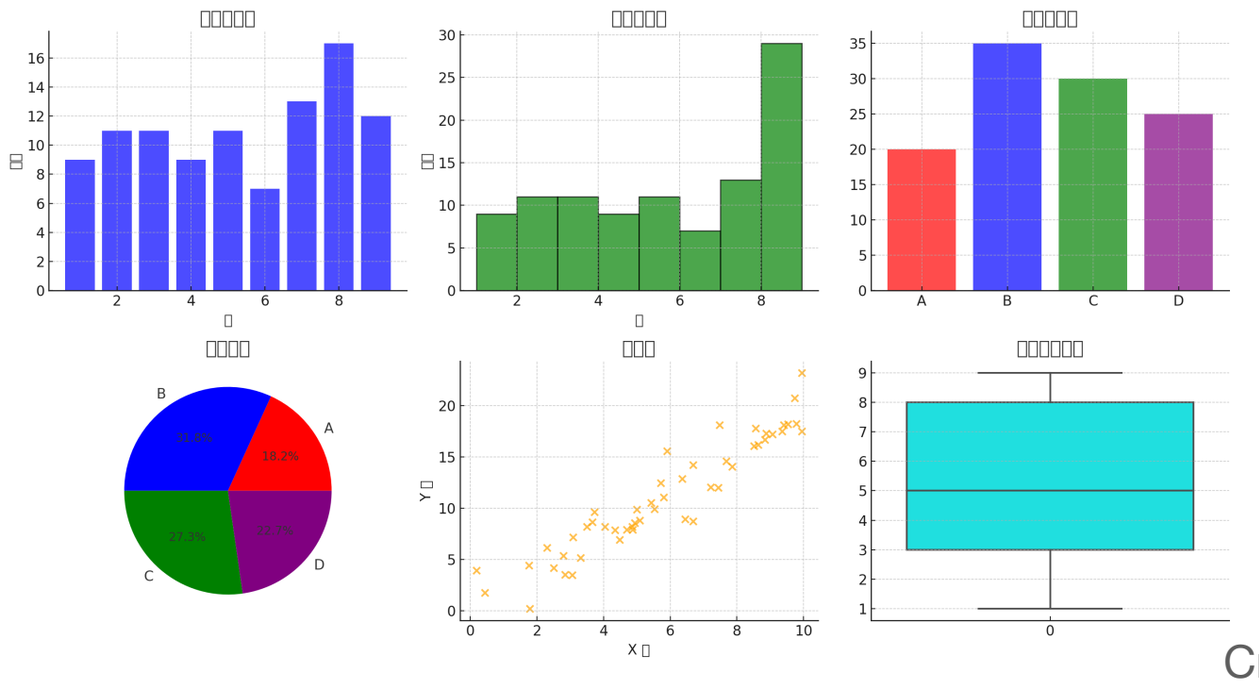


데이터 탐색

데이터 탐색 기초 > 시각적 탐색

- **시각적 탐색**
  - 시각화를 통한 탐색적 자료분석은 전통적 통계차트 및 다이어그램
  - 도수분포표(Frequency table), 히스토그램(histogram), 막대그래프(bar chart), 파이차트(pie chart), 산점도(scatter plot), 줄기 잎 그림(stem-and-leaf diagram), 상자수염그림(box plot)

계급구간(cm)	도수
161.5 이상 165.5 미만	6
165.5 이상 169.5 미만	12
169.5 이상 173.5 미만	18
173.5 이상 177.5 미만	11
177.5 이상 181.5 미만	8
합계	55



Created by ChatGPT

데이터 탐색

고급 데이터 탐색

- **시공간 데이터 탐색**
  - 공간적 정보(데이터)에 시간의 흐름(이력정보 등)이 결합된 다차원 데이터
  - 지리정보 시스템, 위치기반 서비스, 차량 위치추적 서비스 등에 활용

- **다변량 데이터**
  - 변수들 간 인과관계의 규명과 분석, 변수를 축약하거나 분류, 관련된 분석 방법을 적용
  - 다중회귀, 로지스틱 회귀, 분산분석(ANOVA), 다변량 분산분석

데이터 탐색

고급 데이터 탐색

- **비정형 데이터 탐색**
  - 비구조화 데이터, 비구조적데이터 등 미리 정의된 데이터 모델이 없거나 미리 정의된 방식으로 정리되지 않은 정보
  - 전통적인 프로그램을 사용하는 것이 불가능

▸ 데이터 마이닝: 체계적이고 자동적으로 통계적 규칙이나 패턴을 분석하여 가치있는 정보를 추출

- 통계학: 가설 검정, 다변량 분석, 시계열 분석, 선형모델 등
- 데이터베이스: OLAP (온라인 분석처리)
- 인공지능: Self-organizing Map, 신경망

▸ 텍스트 마이닝, 오피니언 마이팅, 웹 마이닝

빅데이터 개론 2

- **빅데이터 탐색**
  - 데이터 전처리
    - 데이터 정제
    - 분석 변수 처리
  - 데이터 탐색
    - 데이터 탐색 기초
    - 고급 데이터 탐색
  - 통계기법 이해
    - 기술통계
    - 추론통계

통계기법 이해

통계기법

- **기술통계 (Descriptive Statistics)**
  - 분석에 필요한 데이터를 요약하여 묘사, 설명하는 통계기법
  - 데이터의 특성을 파악하고 요약(정량화)하는 것이 목표: 평균, 분산, 분포
  - 표본추출(Sampling), 확률분포(Probability Distribution), 표본분포 (Sampling Distribution)

- **추론통계 (Inferential Statistics)**
  - 표본 데이터를 이용해 모집단에 대한 결론 도출
  - 표본을 통해 모집단을 예측 또는 가설 검정하는 것이 목표
  - 추정(estimation), 가설검정(testing hypothesis)

통계기법 이해

기술 통계 > 표본추출

- **표본추출(Sampling)**
  - 모집단 (population): 정보를 얻고자 하는 관심대상의 전체집합
  - 표본 (sample): 모집단의 일부, 원래 집단의 성질을 추측할 수 있는 자료
  - 표본추출 (sampling): 모집단으로부터 표본을 선택하는 과정

- **전수조사와 표본조사:** 대부분 통계조사는 표본조사, 일부의 표본으로 조사분석을 시행하고 모집단 전체의 분석 결과를 사용

- **표본추출 오차:** 모집단의 특성을 과잉대표하거나 최소대표

통계기법 이해

기술 통계 > 표본추출

- **확률 표본추출 기법 (Probability Sampling)**
  - 모집단의 모든 개체가 일정한 확률로 표본으로 선택될 수 있도록 하는 표본추출 방법. 표본이 모집단을 대표할 가능성을 높임.
  - 모든 표본들의 추출확률을 사전에 알 수 있음.
    - 단순무작위추출(simple random): 무작위로 추출
    - 층화추출(stratified): 모집단을 특정기준으로 계층으로 나누고, 각 층에서 무작위로 추출
    - 계통/체계적추출(systematic): 모집단의 요소들을 일정한 간격을 두고 무작위로 추출
    - 군집추출(clustering): 모집단을 여러 개의 집단으로 나눈 후, 군집을 무작위로 선택

통계기법 이해

기술 통계 > 표본추출

- **비확률 표본추출 기법 (Non-Probability Sampling)**
  - 모집단의 모든 개체가 일정한 확률로 표본으로 선택될 수 있도록 하는 표본추출 방법. 표본이 모집단을 대표할 가능성을 높임.
  - 모든 표본들의 추출확률을 사전에 알 수 있음.
    - 간편/편의추출법(convenience): 연구자가 쉽게 접근할 수 있는 방법으로 표본 추출
    - 판단추출법(judgement): 연구자가 대표성이 있다고 판단한 표본을 선택
    - 할당추출법(quota): 모집단의 특정 특성(연령이나 성별)에 따라 할당된 비율대로 표본 선택
    - 눈덩이추출법(snowball): 초기 응답자가 추가 응답자를 소개하는 방식으로 표본을 확장(주로 소수집단 연구에 사용)

통계기법 이해

기술 통계 > 확률분포

- **확률분포(Probability Distribution)**
  - 수치로 대응된 확률변수의 개별 값들이 가지는 확률값의 분포
  - 이산확률분포, 연속확률분포, 확률분포함수
  - 기대값, 분산

통계기법 이해

기술 통계 > 확률분포

- **이산확률분포(Probability Distribution)**
  - 베르누이 분포: 결과가 성공 또는 실패 두 가지로 귀결되는 분포(예-동전)
  - 이항분포: 독립적인 베르누이 시행을 여러 번 반복했을 때, 성공 횟수를 따르는 분포 (예- 동전을 10번 던짐)
  - 다항분포: 세 개 이상의 범주를 가진 확률실험을 여러 번 반복했을 때의 결과 분포 (예- 주사위를 10번 던짐)
  - 포아송분포: 일정한 시간/공간에서 발생하는 사건의 개수(예-콜센터 전화 수)
  - 기하분호: 성공이 나올 때까지 실패한 횟수를 따르는 분포
  - 음이항분포: r번째 성공이 나올 때까지의 실패 횟수를 따르는 분포
  - 초기하분포: 비복원 추출에서 특정 사건이 일어나는 횟수를 따르는 분포

통계기법 이해

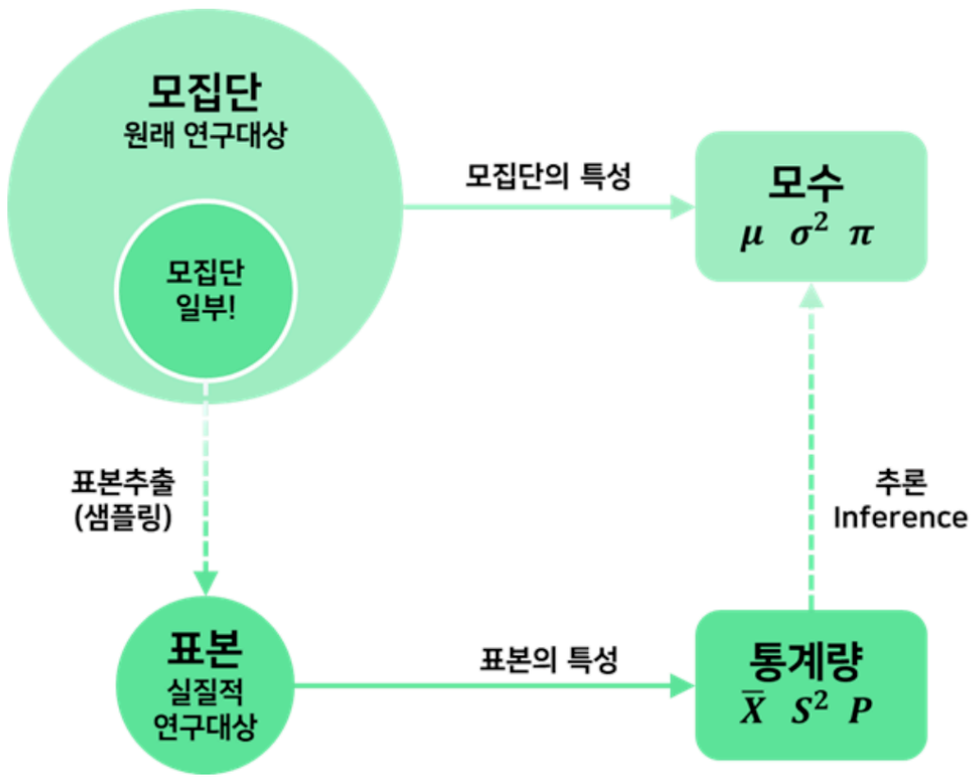
기술 통계 > 확률분포

- **연속확률분포(Probability Distribution)**
  - 연속균등분포: 모든 값이 동일한 확률을 가지는 분포 (예- 0에서 1사이의 실수 랜덤하게 선택)
  - 지수분포: 어떤 사건이 발생할 때까지 걸리는 시간을 나타내는 분포
  - 정규분포: 데이터가 평균을 중심으로 대칭 모양인 분포
  - 표준정규분포: 평균이 0, 분산이 1인 정규분포
  - 감마분포: 지수분포의 확장, 특정 사건이 r번 발생할 때까지 걸리는 시간의 분포
  - 카이제곱분포: 정규분포를 따르는 변수들의 제곱합으로 만들어진 분포, 주로 분산 분석에 사용

통계기법 이해

기술 통계 > 표본분포

- **표본분포(Sampling Distribution)**
  - 크기 n의 확률표본(=무작위로 여러번 뽑은 크기 n의 표본)의 확률변수(=각 표본의 통계값)의 분포



통계기법 이해

기술 통계 > 표본분포

- **표본분포(Sampling Distribution)**
  - 큰수의 법칙(Law of Large Numbers): 표본의 크기 n이 충분히 커지면, 표본 평균은 모집단의 평균에 점점 가까워진다
  - 중심극한정리(Central Limit Theorem): 표본의 크기 n이 충분히 크면, 모집단의 분포가 어떤 형태이든 표본평균의 분포는 정규분포에 가까워진다. = 동일분포를 가지는 분포들의 평균은 그 개수가 많아지면 언제나 정규분포로 수렴

통계기법 이해

추론 통계 > 추정

- **추정(estimation)**
  - 표본을 통해 모집단의 특성을 추측하는 과정
  - 점추정(point estimate)
    - 모집단의 모수에 대한 추정치(평균 또는 표준편차)를 이에 대응하는 통계량으로 추정. 모수를 하나의 값으로 추정
    - 적률방법(Moment Method), 최대우도추정법(Maximum Likelihood Function Method)
  - 구간추정(interval estimate) 또는 신뢰구간(confidence interval)
    - 점추청에 오차의 개념을 도입하여 모집단의 모수가 포함될 가능성이 높은 구간을 특정 신뢰 수준 하에서 제시. 실제 모수가 있다고 예상되는 확률을 기반으로 수행. 신뢰구간/신뢰수준으로 표현

통계기법 이해

추론 통계 > 가설검정

- **가설검정(testing hypothesis)**
  - 모집단에 대한 가설을 설정하고 표본을 분석하여 그 가설의 타당성을 검정하는 과정
  - ▸ 가설검정 절차
  - 가설 설정: 귀무가설(H_0) 대립가설(H_1)
  - 검정통계량 및 표본분포 결정: 표본 데이터에 적용할 검정 방법 선택 (Z-검정, t-검정 등)
  - 기각역(임계값) 설정: 유의수준(α) 설정 (귀무가설 기각할 기준)
  - 검정통계량 계산: 표본에서 검정통계량 계산 후 기각역과 비교
  - 결론: 검정통계량이 기각역에 속하면 H_0 기각 -> H_1 채택