청년을 응원합니다!
K-Digital Training

기초통계 및 경영통계

강사 김경하

고용노동부 | 삼정 KPMG | 현대경제연구원 | i-Growth

과정 소개

1. IT/DX/AIX 기본 역량

- 디지털 이노베이션
- DX, AIX 이해, AI 트랜드
- 파이썬 프로그래밍
- 기초 통계 및 경영 통계
- 생성형 AI 활용
- huggingface 모델 활용
- AI활용 SQL 이해 실습
- 프롬프트 엔지니어링
- streamlit으로 MVP 구현

2. 데이터 분석, 머신러닝

- 데이터 전처리, 시각화
- 탐색적 데이터 분석
- 데이터 수집(정적/동적/API)
- 머신러닝 학습방법
  - 지도학습 – 회귀, 분류
  - 비지도학습 – 군집
  - 머신러닝 모델 평가
- 딥러닝 이해, 기본 학습
  - CNN, 얼굴인식, 객체 탐지

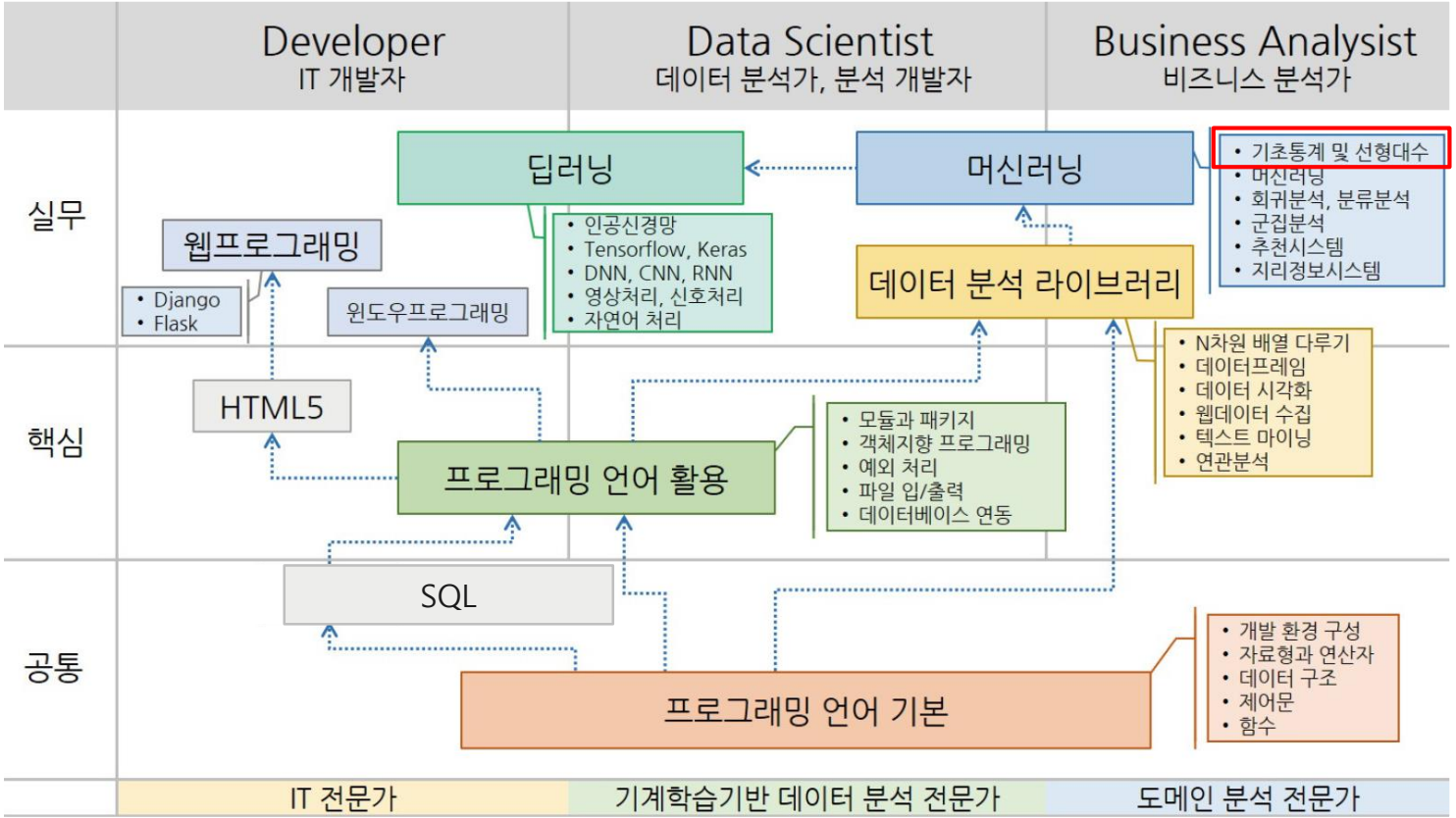
3. 생성형 AI, RAG 구현

- LangChain 기본
- RAG 이해 및 활용
- AI Agent, MCP 활용
- AI 모델 서빙, 백엔드 개발
- 프로젝트 관리 이해
- 프로젝트 기획, 설계
- WBS 이해 및 작성
- 도커기반, AWS 클라우드 배포

MVP(Minimum Viable Product; 최소 기능 제품)

파이썬 언어 활용 학습 로드맵

파이썬을 이용한 머신러닝 데이터 분석



머신러닝/딥러닝을 학습 하기 위한 확률, 통계 및 수학적 개념의 기본을 익힘

출처: 한국소프트웨어산업협회

강의 진행 계획(10/14~20)

- 통계 기본 익히기, 데이터 수집 방법 익히기, 파이썬 고급

시간	내용
1 ~ 4교시	기본 이론, 실습
5~6교시	최신 AI 적용 사례 조사 및 발표
6~7교시	기본 이론, 실습
8교시	내용 정리, 마무리

기초 통계 with 데이터분석, 머신러닝

학습 목표

- 탐색적 데이터 분석(EDA)을 이해한다.
- 통계학 용어에 익숙해 진다.
- 통계학 용어와 통계 모델링에 대해 안다.
- 대표통계, 산포통계, 분포통계 개념을 알다
- 기술 통계학과 추론 통계학의 핵심 개념을 이해한다

통계 용어와 통계모델 이해

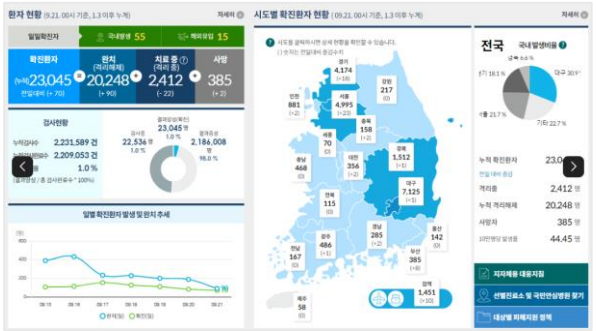
일화적 증거

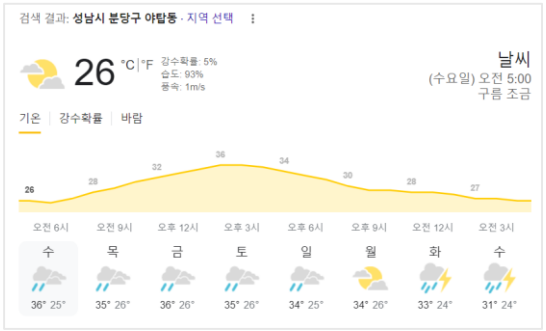
- 요즘 애들 대부분 핸드폰을 하루에 5시간 이상한다.
- 옆집에도 하루에 5시간 이상하고, 동생내도 아이들이 하루에 5시간 이상 하고, 우리 아들도 핸드폰 하기 위해 공부하는 모양으로 핸드폰 시간이 6시간이 넘는다.
- 그러니 요즘애들은 핸드폰을 하루에 5시간 이상할 것 같다.

- 일화적 증거의 신빙성이 부족한 이유
  - 적은 관측치
  - 선택편의
  - 확증편의(편향)
  - 부정확함

통계학의 기원과 활용

- 통계학의 기원
  - 라틴어의 'status' = '국가에 대한 모든 사실' 로부터 파생되었다는 설
  - 이태리어의 'state' = '국가'로 부터 파생되었다는 설
- 통계의 활용
  - 국가 행정에서 시작, 고대의 징세와 징병을 목적으로 사용됨
  - 현대는 일기예보나 실시간 교통정보, 주식시장분석 등 많은 곳에 사용
  - 컴퓨터의 발달은 통계학의 활용성을 넓히는데 기여함.





통계학과 경영통계

- 통계학 정의
  - 통계학은 **자료를 의사결정에 도움이 되는 의미 있는 정보**로 전환하여 **사회현상이나 자연현상을 수치로 요약**하는 학문

- 경영통계
  - **조직이나 사람, 회사**라는 것이 어떻게 구성되어 있고, 구성원들은 어떻게 생각을 하고, **고객의 관점** 등 다양한 과점에 대해서 **과학적으로 살펴**보는 것
  - 의사결정, 계량경제, 재무분석 등에 활용

통계적 접근법

- 자료수집(Data collection) : 대표성을 띠는 자료 수집
- 기술통계(Descriptive statistics) : 데이터 통계량 요약, 정리, 시각화
- 탐색적 데이터 분석(Exploratory data analysis)
  - 데이터의 패턴, 차이점, 특이점 찾기
- 추정(Estimation) : 사용한 데이터(표본)를 토대로 모집단 특징 추정
- 가설검정(Hypothesis testing) : 데이터셋에서 두 그룹 간의 차이가 명백한 것인지 혹은 우연한 사건인지 평가

탐색적 데이터 분석(EDA; Exploratory Data Analysis)

그래프를 통한 **시각화와 통계 분석** 등을
바탕으로 수집한 데이터를 **다양한 각도에서**
관찰하고 이해하는 방법

통계학이란?

수치 데이터의 **수집, 분석, 해석, 표현** 등을 다루는 수학의 한 분야

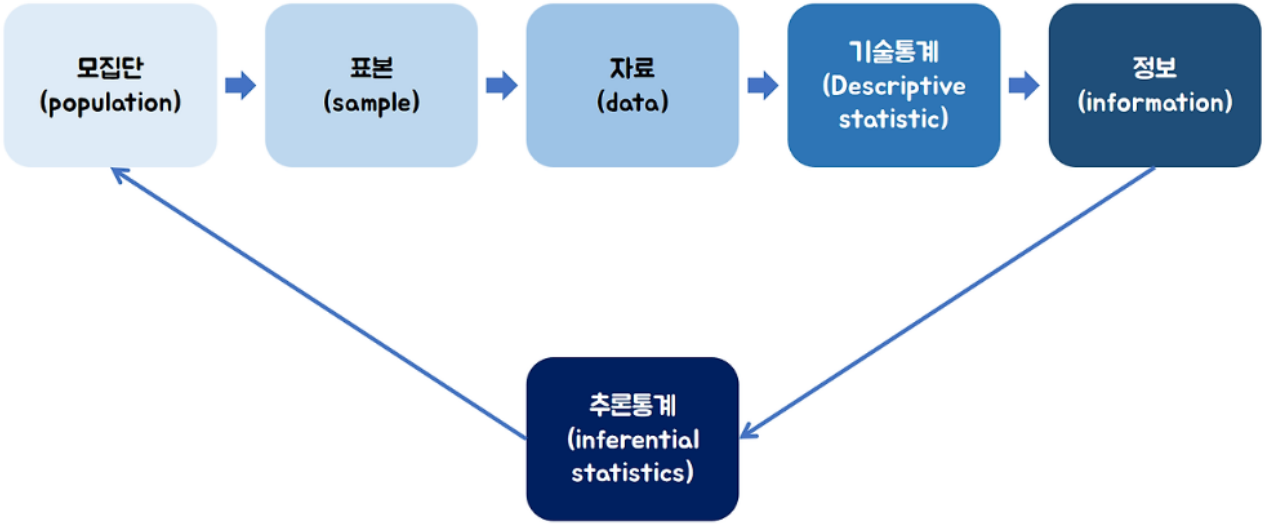
- **기술 통계학**(Descriptive Statistics)
  - 수집된 데이터의 주요 특징을 요약하고 설명함.
  - 복잡하고 많은 양의 데이터를 간결하고 이해하기 쉽게 전반적인 경향, 분포, 형태 등을 파악하는 데 초점
  - 일상생활에서 쓰는 통계들
- **추론 통계학**(Inferential Statistics)
  - **일부 자료를 수집(표본)**해서 **전체 모집합에 대한 결론을 유추해냄**
  - **가설검정, 수치의 특징 계산, 데이터 간의 상관관계** 등을 활용함
- **통계 모델링**(Statistical Modeling)
  - 데이터에 통계학을 적용해 **변수의 유의성을 분석함**으로써 **데이터의 숨겨진 특징**을 찾아내는 것

통계학 – 기술통계

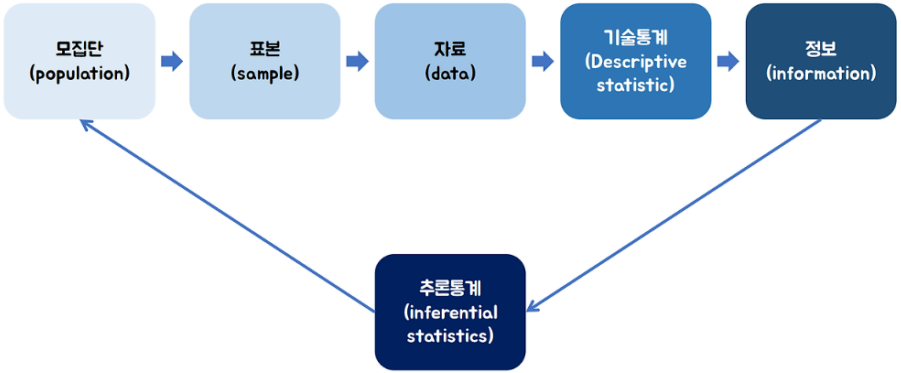
- 수치형 데이터 (Numerical Data)
  - 숫자로 표현되는 데이터를 의미함.
  - 연속형(측정): 키, 나이, 가격 등
  - 이산형(세기): 자동차 수, 학생 수 등
  - 평균, 표준편차와 같은 자료 요약
- 범주형 데이터
  - 명확하게 구분되는 범주 또는 그룹으로 나눌 수 있는 데이터
  - 명목형(순서없음): 성별, 혈액형, 거주도시
  - 순서형(순서있음): 학력, 만족도
  - 빈도, 백분율과 같은 자료 요약

통계학 – 추론통계

- 부분의 데이터를 이용해서 전체를 설명하는 것
  - 대통령 지지율의 경우에 a 후보와 b후보의 지지율을 이야기 하면서 표본오차를 설명함.



통계학 – 추론통계



구분	설명
모집단	전체를 의미함. 모(母)를 사용
표본	샘플, 전체에서 일부분의 데이터 사용
자료	표본에서 나온 결과값의 정보, 일정수의 사람들의 찬반의 의견을 밝히면 이게 바로 자료
기술통계	"전체 200명 중 찬성이 45%, 반대가 40%, 나머지 무응답 15%, 찬성이 더 많다"로 설명함
추론통계	궁금한 것은 전체를 알고 싶어서 샘플링 한 것, 부분의 통계로 전체를 추론하는 것

통계학 – 통계모델

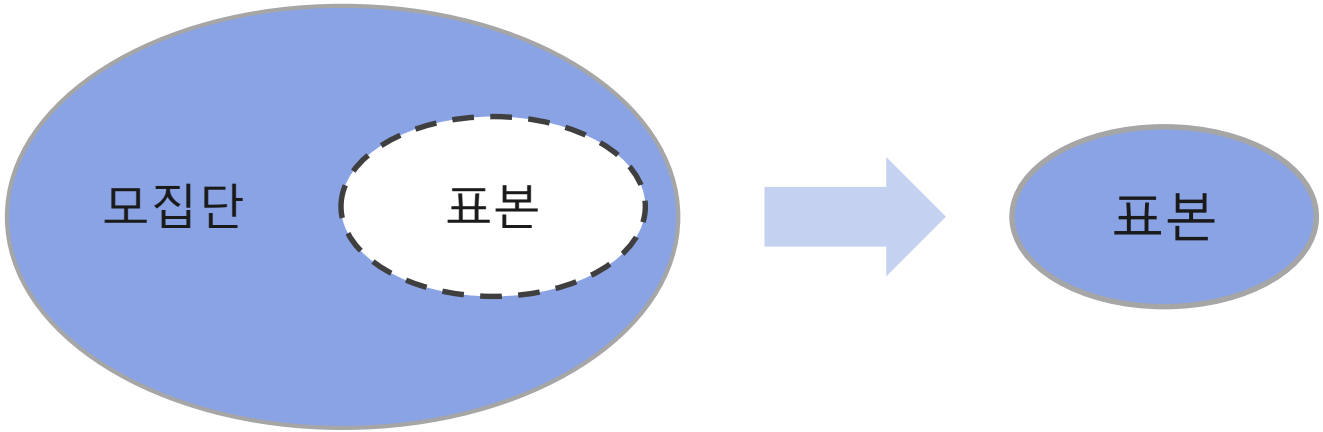
- **통계모델**
  - 수학적 모델 : 변수들로 이뤄진 **수학식을 계산**해 **실제 값을 추정**하는 방법
  - 통계 모델을 이루는 여러 가정은 확률 분포를 따름 : **이산 확률 분포**와 **연속 확률 분포**
  - **모든 변수가 만족해야 하는 기본 가정으로 시작**하며, 이 조건이 만족할 때만 **모델의 성능이 통계학적으로 의미**를 가짐

- 머신러닝 분야의 발전 방향
  - 수학 -> 통계학 -> 컴퓨터 과학 -> 머신러닝(딥러닝)

통계학 – 통계모델

모집단 모든 관측값 또는 분석 대상의 전체 데이터를 의미

표본 모집단의 부분 집합으로, 분석 대상 중인 전체 데이터의 일부분



통계학 – 통계모델

- 모수 : **모집단의 특징**을 나타내는 **수치값**
- 통계량
  - **표본의 특징을 나타내는 수치값**으로, **모수 추정**을 위해
  - 사용통계량을 계산하는 **기초통계분석(기술통계분석)**을 바탕으로 **확률 분포**를 간단하게 확인 가능
  - **평균, 중앙값, 최빈값, 분산** 등과 같은 데이터를 대표하는 값

통계량이 실제 모집단을 **대표하는 값**이 될 때,
통계적 유의성을 확보할 수 있음

통계학 – 통계모델

통계모델 종류

구분	종류
회귀모델 (Regression)	- 선형 회귀 (Linear Regression): 연속형 종속 변수와 하나 이상의 독립 변수 간의 선형 관계를 모델링함 - 로지스틱 회귀 (Logistic Regression): 범주형 종속 변수(특히 이진 분류)의 확률예측 - 다항 회귀 (Polynomial Regression): 종속 변수와 독립 변수 간의 비선형 관계를 모델링함
분류모델 (Classification)	- 의사 결정 트리 (Decision Trees): 데이터를 기반으로 의사 결정 규칙을 학습하여 범주를 예측함 - 서포트 벡터 머신 (Support Vector Machines, SVM): 데이터를 가장 잘 분리하는 초평면을 찾아 범주를 분류함
시계열모델 (Time Series)	- ARIMA(Autoregressive Integrated Moving Average): 시간 순서대로 배열된 데이터의 패턴을 분석하고 예측함 - 지수 평활법 (Exponential Smoothing): 과거 값에 가중치를 부여하여 미래 값을 예측
군집화모델 (Clustering)	K-평균 (K-Means): 유사한 데이터 포인트를 여러 개의 그룹(클러스터)으로 나눔

통계학이란?

수치 데이터의 **수집, 분석, 해석, 표현** 등을 다루는 수학의 한 분야

- **기술 통계학**(Descriptive Statistics)
  - 수집된 데이터의 주요 특징을 요약하고 설명함.
  - 복잡하고 많은 양의 데이터를 간결하고 이해하기 쉽게 전반적인 경향, 분포, 형태 등을 파악하는 데 초점
  - 일상생활에서 쓰는 통계들
- **추론 통계학**(Inferential Statistics)
  - **일부 자료를 수집(표본)**해서 **전체 모집합에 대한 결론을 유추해냄**
  - **가설검정, 수치의 특징 계산, 데이터 간의 상관관계** 등을 통해 이뤄짐
- **통계 모델링**(Statistical Modeling)
  - 데이터에 통계학을 적용해 **변수의 유의성을 분석함**으로써 **데이터의 숨겨진 특징**을 찾아내는 것

기초 통계 분석

기초 통계 분석

- 통계량의 분류

구분	내용	예시
대표 통계량	데이터의 **중심 및 집중경향**을 나타내는 통계량	평균, 최빈값 등
산포 통계량	데이터의 **퍼진 정도**를 나타내는 통계량	분산, 표준편차, 범위
분포 통계량	데이터의 **위치 정보 및 모양**을 나타내는 통계량	왜도, 첨도, 사분위수, 최대값, 최소값

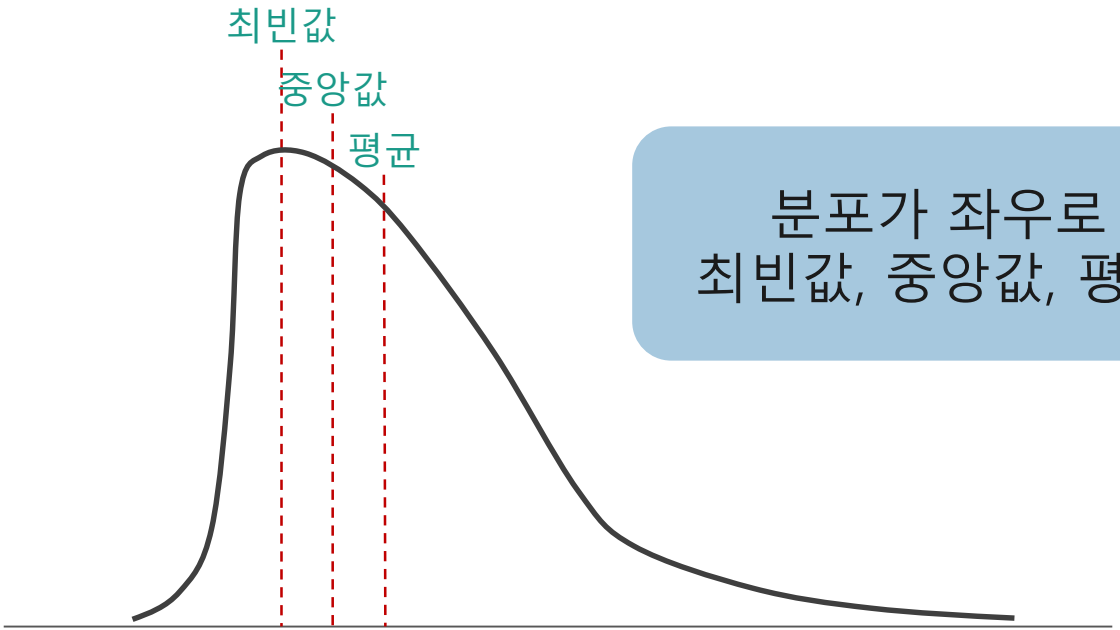
대표 통계량

- 통계량의 분류

구분	내용	예시
대표 통계량	데이터의 **중심 및 집중경향**을 나타내는 통계량	평균, 최빈값, 중앙값등
산포 통계량	데이터의 **퍼진 정도**를 나타내는 통계량	분산, 표준편차, 범위
분포 통계량	데이터의 **위치 정보 및 모양**을 나타내는 통계량	왜도, 첨도, 사분위수, 최대값, 최소값

대표 통계량

데이터의 대푯값(Representative value)
: 중심 경향(central tendency) 값이라고도 함.



분포가 좌우로 치우친 경우
최빈값, 중앙값, 평균이 모두 다름

참고 : https://brunch.co.kr/@data/13

대표 통계량 – 평균

- **산술평균** : 가장 널리 사용되는 평균으로 연속형 변수 사용
  - 이진 변수 대한 산술 평균은 1의 비율과 같음
  - 다른 관측치에 비해 매우 크거나 작은 값에 크게 영향을 받음
- **조화평균** : 비율 및 변화율 등에 대한 평균을 계산할 때 사용
  - 데이터의 역수의 산술평균의 역수
- **절사평균** : 데이터에서 α~1- α 의 범위에 속하는 데이터에 대해서만 평균을 낸 것
  - 매우 크거나 작은 값에 의한 영향을 줄이기 위해 고안됨

대표 통계량 – 평균

- 파이썬을 이용한 평균 계산

구분	구현 함수
산술 평균	• numpy.mean(x) • numpy.array(x).mean() • Series(x).mean()
조화 평균	• len(x) / numpy.sum(1/x) • scipy.stats.hmean(x)
절사 평균	• scipy.stats.trim_mean(x, proportiontocut) - proportiontocut: 절단한 비율

대표 통계량[실습]

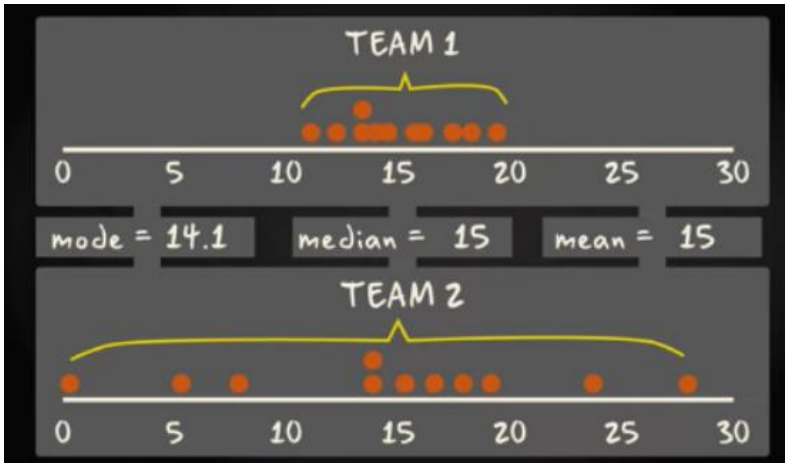
- 최빈값
  - 한 변수가 **가장 많이 취한 값**을 의미함.
  - **범주형 변수**에 대해서만 적용
  - 파이썬을 이용한 최빈값 계산

[실습1]통계_대표 통계량

데이터 분석 및 머신러닝을 위한 가상환경 만들기
- 가상환경 이름 : eda_env
- python 버전 : 3.12
- 가상환경 활성화
- jupyter note북을 사용하기 위한 라이브러리 설치 및 설정
  pip install ipykernel
  python -m ipykernel install --user --name 가상환경이름 --display-name "커널출력이름"

대표 통계량의 문제점

- Team1과 Team2는 분명히 다른 값을 갖고 있지만 평균값과 중위 값, 최빈값은 동일하게 나타남
- 대푯값 만으로는 데이터를 설명하는데 한계가 있음



출처 : cousera.org

데이터의 변화량을 나타내는 것이 필요함.

산포 통계량

- 통계량의 분류

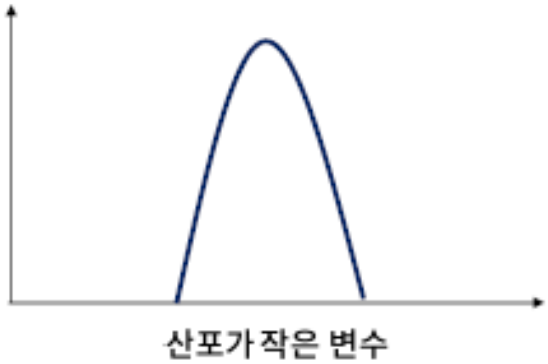
구분	내용	예시
대표 통계량	데이터의 **중심 및 집중경향**을 나타내는 통계량	평균, 중앙값, 최빈값 등
산포 통계량	데이터의 **퍼진 정도**를 나타내는 통계량	분산, 표준편차, 범위
분포 통계량	데이터의 **위치 정보 및 모양**을 나타내는 통계량	왜도, 첨도, 사분위수, 최대값, 최소값

산포 통계량

- 분산(variance) : 평균과의 거리를 제곱한 값의 평균
- 표준편차(standard deviation) : 분산의 제곱근
- 범위(range) : 최대값과 최소값의 차이

산포 통계량

- 산포 : 데이터가 **얼마나 퍼져 있는지**를 의미함.



- 즉, 산포 통계량이란 데이터의 산포를 나타내는 통계량

산포 통계량-분산, 표준편차

- 편차 : 한 샘플이 평균으로부터 떨어진 거리

$$x_i - \mu \ (x_i : i\text{번째 관측치}, \mu : \text{평균})$$

- 분산(variance)
  - 편차의 제곱평균, 편차의 합은 항상 0이 되기 때문에, 제곱을 사용
  - 샘플 데이터는 n-1로 나눔
  - 자유도가 0(모분산)이면, n으로 나눔

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

x_i : i번째 관측치
μ : 평균
n : 관측치의 개수

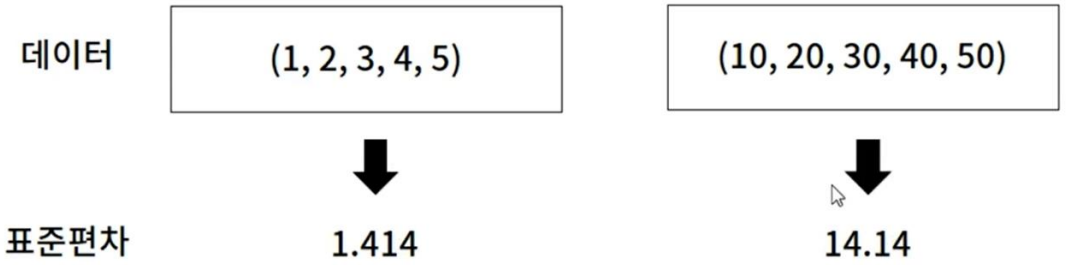
- 표준편차(standard deviation)
  - 분산의 제곱근

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

x_i : i번째 관측치
μ : 평균
n : 관측치의 개수

산포 통계량-변동계수

- 분산과 표준편차 모두 값의 스케일에 크게 영향을 받아, 상대적인 산포를 보여주는데 부적합함.



- 따라서 **변수를 스케일링한 뒤, 분산 혹은 표준편차**를 구해야 함.
- 만약 모든 데이터가 양수인 경우에는 변동계수(상대 표준편차)를 사용할 수 있음
  - 변동계수 = 표준편차 / 평균

파이썬을 이용한 분산, 표준편차, 변동계수 계산

구분	구현 함수	비고
분산	• numpy.var(x, ddof) • numpy.array(x).var(ddof) • Series(x).var(ddof)	ddof : 자유도
표준편차	• numpy.std(x, ddof) • numpy.array(x, ddof).std() • Series.std(x, ddof)	
변동계수 계산	• numpy.std(x, ddof) / numpy.mean(x) • scipy.stats.variation(x)	

- [실습2]통계_산포 통계량

산포 통계량 - 스케일링(Scaling)

- 둘 이상의 변수의 값을 상대적으로 비교할 때 사용
- 예) 국어 점수 = 85점인 학생, 수학 점수 = 75점인 학생 누가 더 잘했나?
  - 국어 점수 변수와 수학 점수 변수의 분포가 다르기 때문에 정확히 비교가 힘듬
  - 국어 점수 평균 = 90점, 수학 점수 평균 = 35점
  - 상대적으로 봤을 때 수학점수 75점이 더 잘한 것임.

산포 통계량 - 스케일링(Scaling)

- 상대적으로 비교하기 위해 각 데이터에 있는 값을 상대적인 값을 갖도록 변환함

$$\frac{x-\mu}{\sigma}$$

Standard Scaling

$$\frac{x-\min(x)}{\max(x)-\min(x)}$$

Min-max Scaling

- 스케일링은 변수간 비교 및 머신러닝에서 피처간의 데이터 수준을 맞춰주기 위해 많이 사용됨

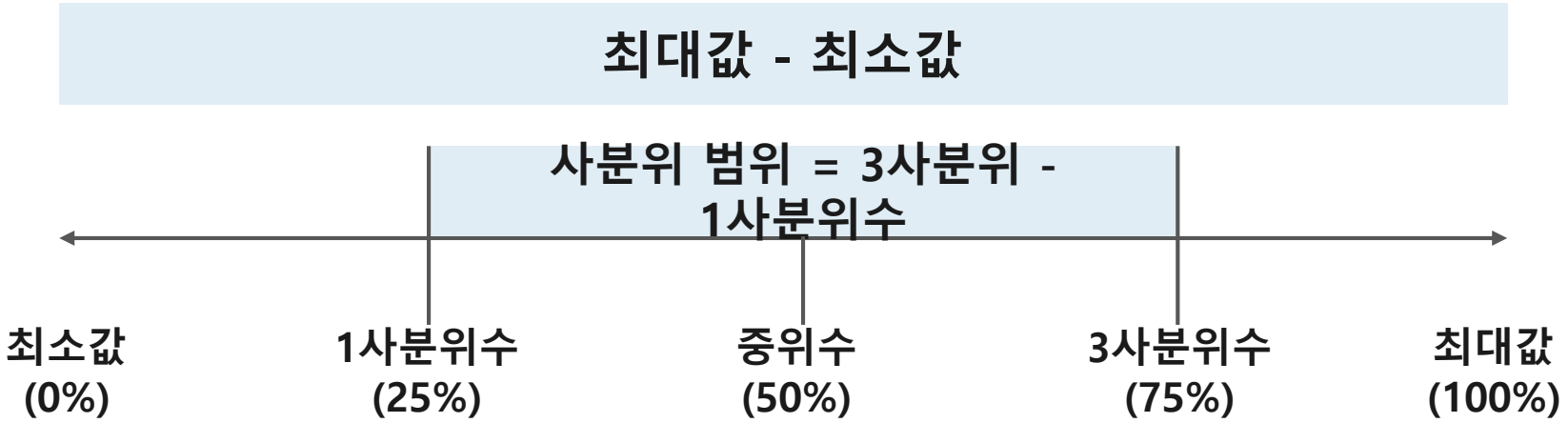
산포 통계량 - 스케일링(Scaling)

구분	구현 함수
Standard Scaling	• (x-x.mean()) / x.std() # x: ndarray • **sklearn.preprocessing.StandardScaler**
Min-max Scaling	• x-x.min() / (x.max() – x.min()) # x: ndarray • **sklearn.preprocessing.MinMaxScaler**

- [실습2]통계_산포 통계량

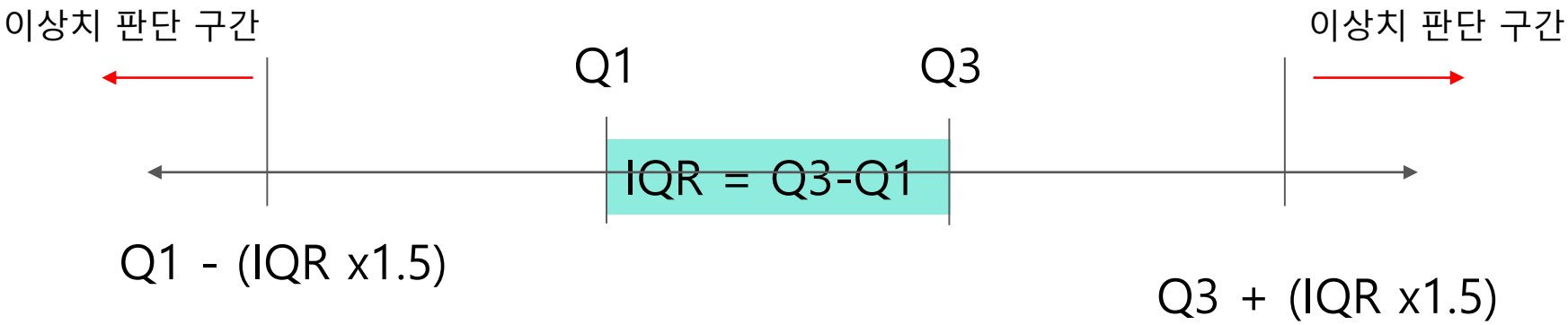
범위와 사분위 범위

- 범위와 사분위 범위는 산포를 나타내는 직관적인 지표
- IQR(Interquartile Range)라고하며, 이상치 탐색 시 활용함



IQR Rule

- 변수별로 IQR 규칙을 만족하지 않는 샘플들을 판단하여 삭제하는 방법



파이썬을 이용한 범위 및 사분위 범위 계산

구분	구현 함수
범위	• numpy.ptp(x) • numpy.max(x) – numpy.min(x)
사분위 범위	• numpy.quantile(x, 0.75) – numpy.quantile(x, 0.25) • scipy.stats.iqr(x)

- [실습2]통계_산포 통계량

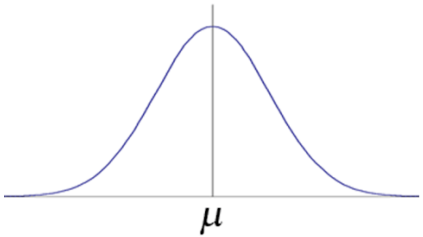
기초 통계 분석

- 통계량의 분류

구분	내용	예시
대표 통계량	데이터의 **중심 및 집중경향**을 나타내는 통계량	평균, 최빈값 등
산포 통계량	데이터의 **퍼진 정도**를 나타내는 통계량	분산, 표준편차, 범위
분포 통계량	데이터의 **위치 정보 및 모양**을 나타내는 통계량	사분위수, 최대값, 최소값, 왜도, 첨도

분포 통계학

- 최댓값(maximum), 최솟값(minimum)
- 사분위수(quartiles) : 자료를 4등분한 값
- 왜도(skewness), 첨도(kurtosis) : 데이터 분포의 **대칭 정도**와 데이터의 **이상치(outliers) 존재 여부**를 나타냄
- 정규분포(normal distribution)

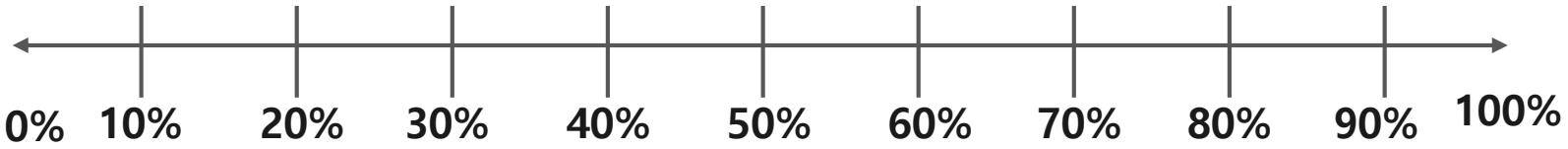


분포 통계량 – 최댓값과 최솟값

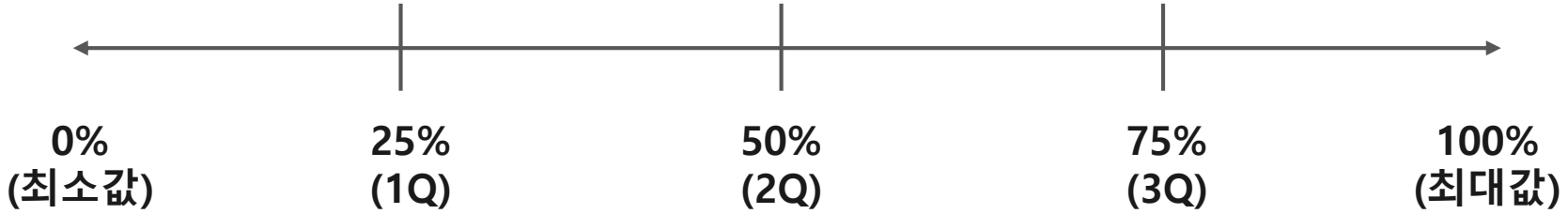
- 범위(range)
  임의의 값을 가지는
  데이터 집합에서
  최댓값과 최솟값의 차이
  - Team1 : 8.5
  - Team2 : 27.7

분포 통계량 – 백분위수와 사분위수

- **백분위수(percentile):** 데이터를 크기 순서대로 **오름차순 정렬**했을 때, **백분율**로 나타낸 특정 위치의 값을 의미함

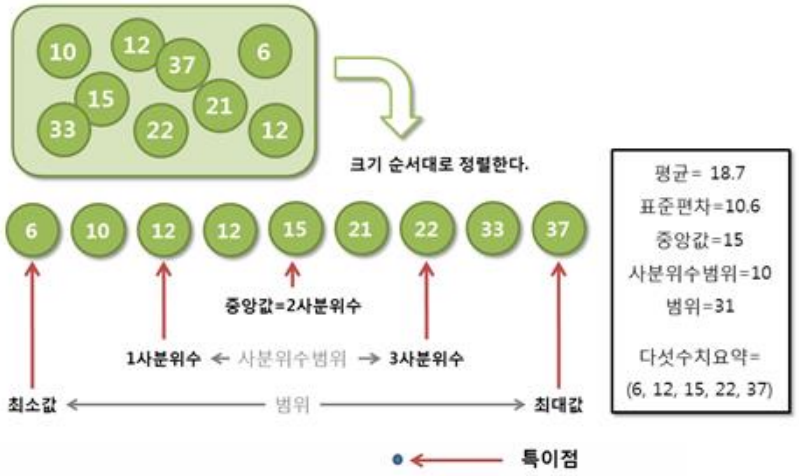


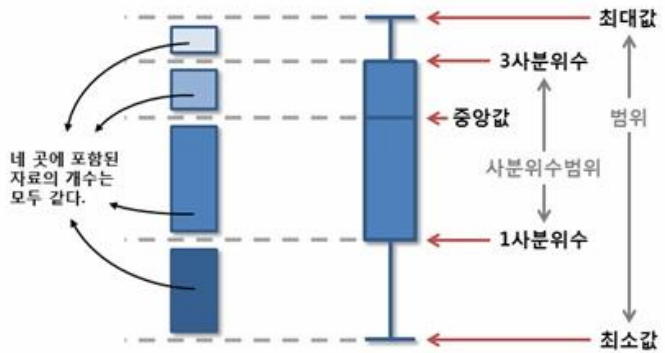
- **사분위수(quantile)**: 데이터를 크기 순서대로 **오름차순 정렬**했을 때, **4등분한 위치**의 값을 의미함



분포 통계량 - 사분위수

- 사분위수
  : 데이터를 4등분한 값
  - 25%값:1사분위수(Q1),
  - 50%값:2사분위수(Q2),
  - 75%값:3사분위수(Q3)
  - IQR : Interquartile Range.
    (Q3 – Q1)

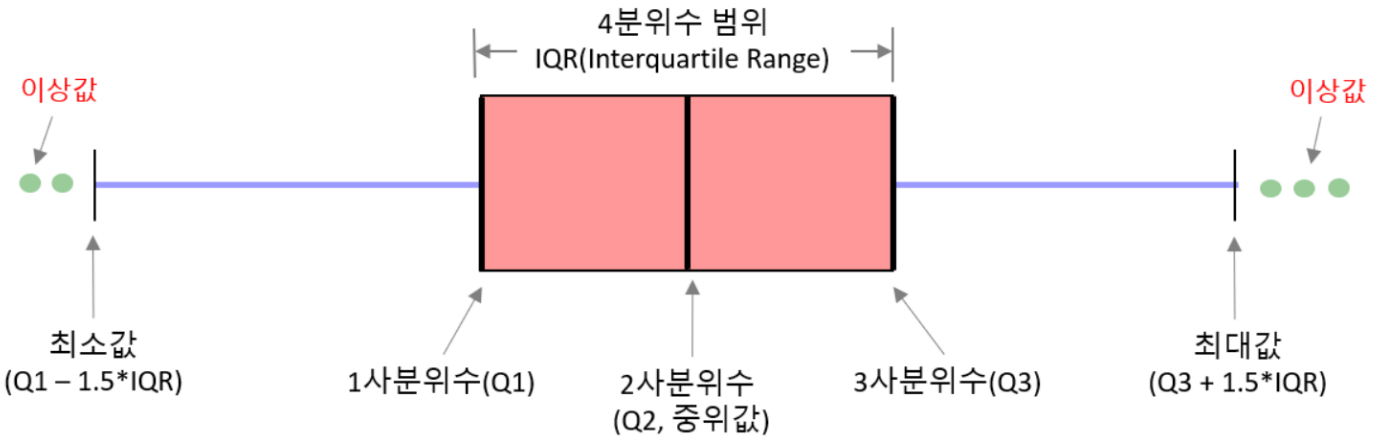




참고: https://t1.daumcdn.net/cfile/tistory/2223A24654298E1415

분포 통계량 - 사분위수

- Box plot에서 활용
  - 박스 플롯은 관측값의 대략적 분포와 개별적 이상치를 파악할 수 있는 시각화 차트,
  - 한 공간에서 여러 개의 관측값 그룹 시각화

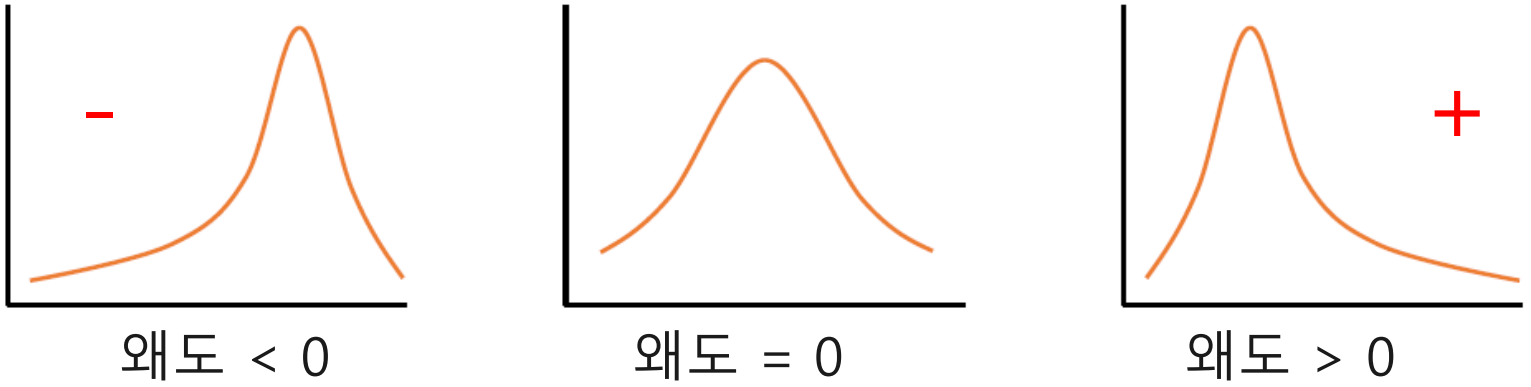


분포 통계량 – 백분위수와 사분위수

구분	구현 함수	비고
백분위수	• numpy.percentile(x, q)	q: 위치(0~100)
사분위수	• numpy.quantile(x, q)	q: 위치(0~1)

분포 통계량 – 왜도

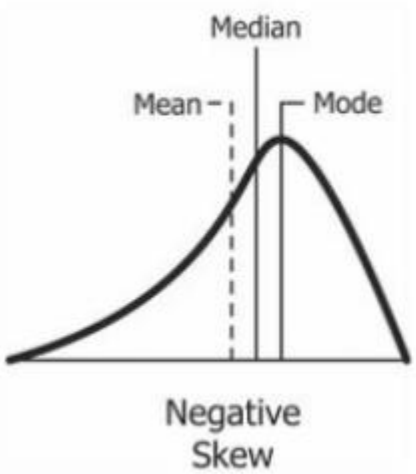
- 왜도(skewness) : 분포의 비대칭도를 나타내는 통계량
- 왜도가 음수면 오른쪽으로 치우친 것을 의미하며,
  양수면 왼쪽으로 치우침을 의미함



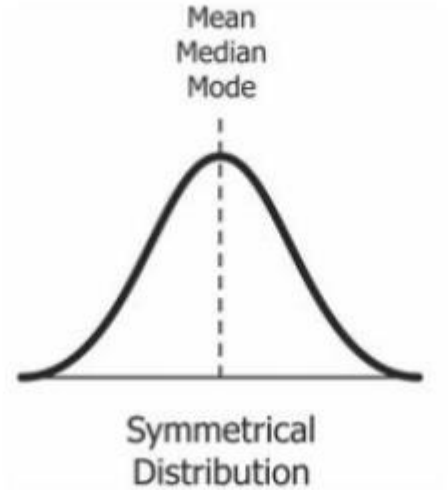
- 일반적으로 왜도의 절대값이 1.5이상이면 치우쳤다고 봄
- 왜도가 치우치면 모델의 일반화가 어려울 수 있음

분포 통계량 – 왜도

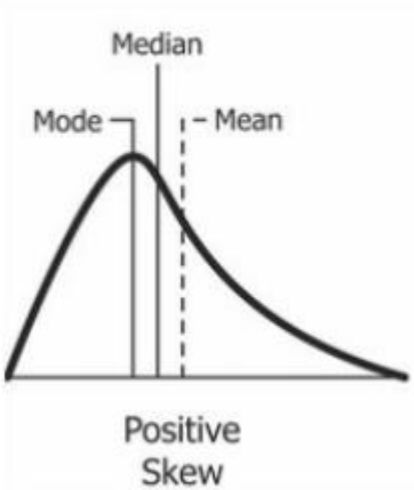
- 왜도
  - 분포가 정규분포에 비해 얼마나 비대칭인지를 나타내는 지표



왜도< 0



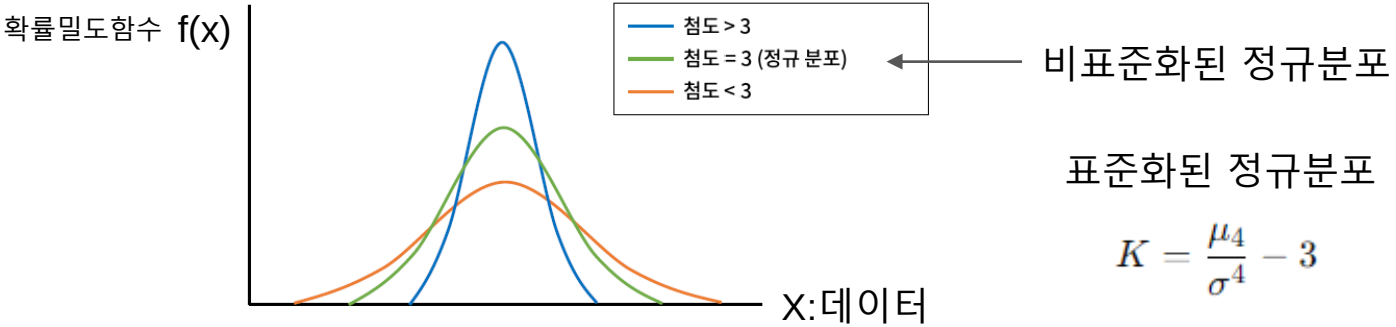
왜도= 0



왜도> 0

분포 통계량 – 첨도

- 첨도(kurtosis)
  - 데이터 분포의 모양을 나타내는 지표
  - 뽀족함(Peak) : 데이터가 얼마나 평균 근처에 몰려 있는지 판단
  - 꼬리(Tail) : 극단적인 값이 얼마나 자주 나타나는지 판단
- 해석
  - 첨도 높음: 아주 드물게 대박 또는 쪽박 같은 극단값이 나올 수 있음
  - 첨도 낮음: 안정적이며 예측 가능한 경향성이 있음



비표준화된 정규분포

표준화된 정규분포

$$K = \frac{\mu_4}{\sigma^4} - 3$$

분포 통계량 – 첨도

구분	구현 함수
왜도	• scipy.stats.skew(x) • Scries(x).skew()
첨도	• scipy.stats.kurtosis(x) • Series(x).kurtosis()

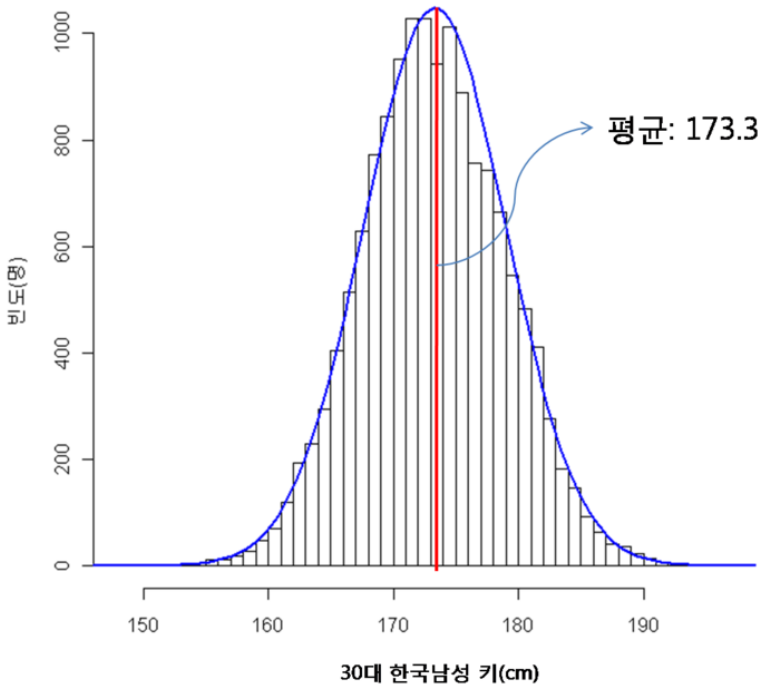
- [실습3]통계_분포 통계량

문제해결

- 왜도를 보는 이유는?
- 첨도를 보는 이유는?
- 경영환경 데이터에서 적용한 예시 조사해서 발표

분포 통계량 - 정규분포

- 정규분포(normal distribution)
  : 평균을 중심으로 해서 사건이 중앙에 가장 많이 분포하고 양끝으로 갈수록 희박하게 분포하며, 평균을 축으로 그래프의 양쪽이 정확히 겹쳐짐



정리

기초 통계 분석 방법

- 통계량의 분류

구분	내용	예시
대표 통계량	데이터의 **중심 및 집중경향**을 나타내는 통계량	평균, 최빈값 등
산포 통계량	데이터의 **퍼진 정도**를 나타내는 통계량	분산, 표준편차, 범위
분포 통계량	데이터의 **위치 정보 및 모양**을 나타내는 통계량	왜도, 첨도, 사분위수, 최대값, 최소값

- 분산, 표준편차, 범위, 사분위수, IQR 구하기

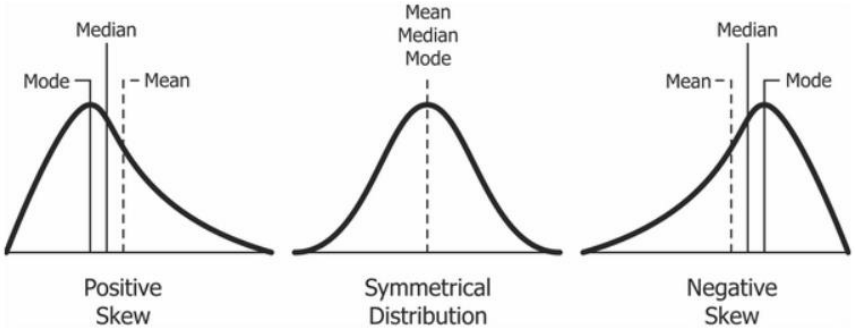
```python
import numpy as np
from statistics import variance, stdev

np.random.seed(0)

points = np.random.randint(0, 100, 20)
var = variance(points); print("분산: ", var)
std = stdev(points); print("표준편차: ", np.round(std, 2))
range = np.max(points) - np.min(points); print("범위: ", range)
print("사분위수:")
for val in [0, 25, 50, 75, 100]:
    quantile = np.percentile(points, val)
    print("{}% => {}".format(val, quantile))

q1, q3 = np.percentile(points, [25, 75])
print("IQR: ", q3 - q1)
```

```
분산:  662
표준편차:  25.73
범위:  79
사분위수:
0% => 9.0
25% => 42.75
50% => 64.5
75% => 84.0
100% => 88.0
IQR:  41.25
```

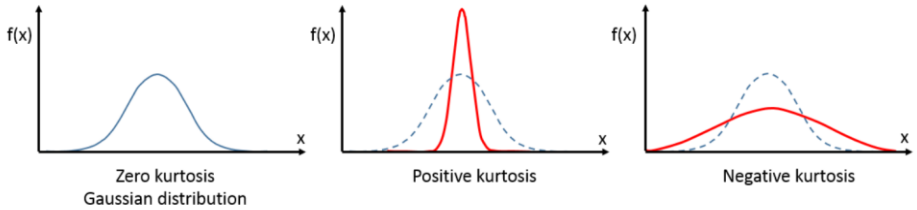
머신러닝을 위한 통계학 핵심 개념

분포

- 왜도:데이터 분포의 대칭 정도를 나타냄
  - 양의 왜도 : 오른쪽 꼬리가 김
  - 음의 왜도 : 왼쪽 꼬리가 김



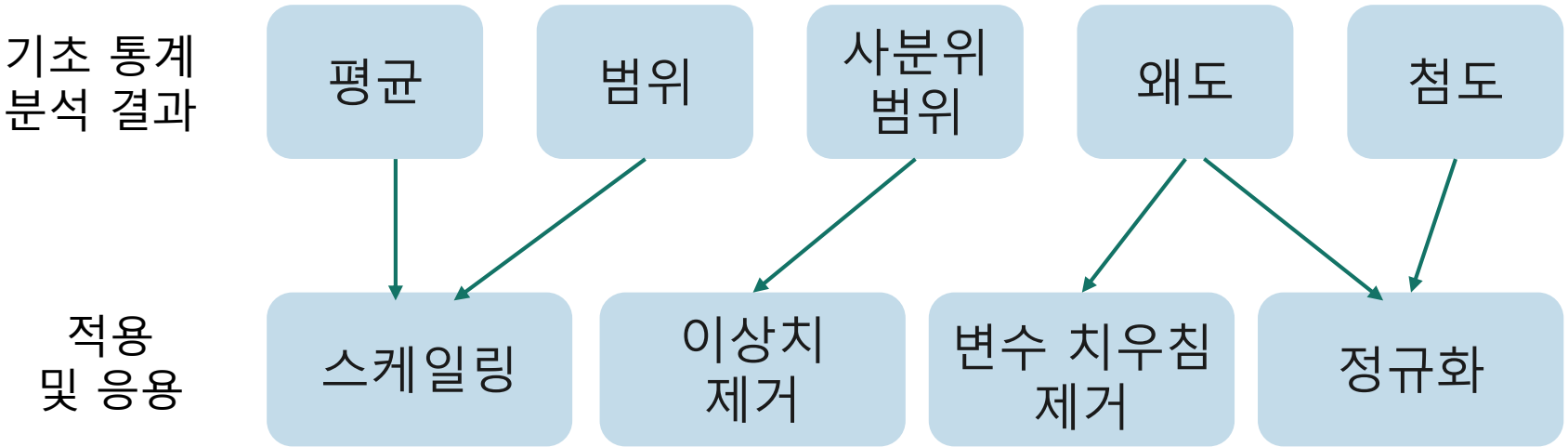
- 첨도:데이터의 이상치(outliers) 존재여부를 나타냄
  - 양의 첨도 : 뾰족한 분포
  - 음의 첨도 : 납작한 분포



참고: https://brunch.co.kr/@data/13

기초 통계 분석 활용 방법

- 머신러닝에서 각 **변수를 이해**하고, 특별한 분포 **문제가 없는지 확인**하기 위해 기초 통계 분석을 수행함




감사합니다