통계용어

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

통계학이란?

통계학

- 주어진 데이터를 수집, 분석, 해석하여 결론을 도출하는 학문
- 특정 현상에 대한 패턴을 파악, 예측할 수 있다
- 과학, 경제, 사회 의료 등 다양한 분야에서 활용된다

통계 용어 정리

데이터의 종류

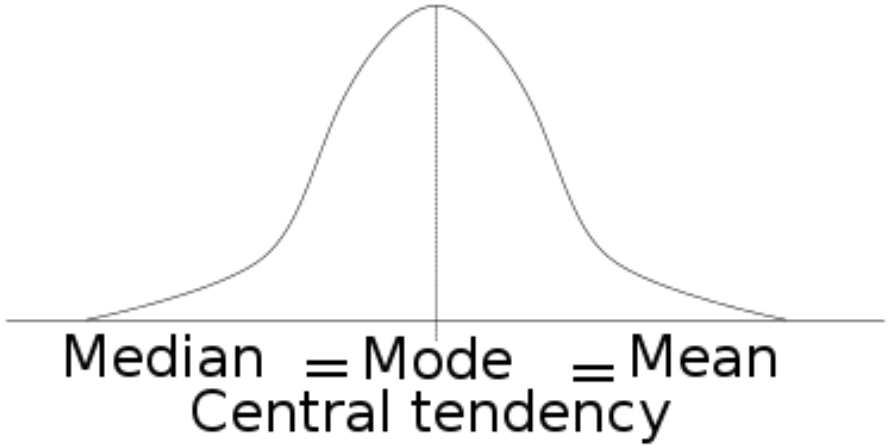
- 양적 데이터
  - 수치로 표현되는 데이터

  Ex) 온도, 키, 몸무게

- 질적 데이터
  - 명목형 데이터
    - 숫자로 표현되지 않는 데이터
      - Ex) 성별, 혈액형, 지역
  - 순서형 데이터
    - 숫자로 표현이 되지만 등간이나 비율의 의미는 없는 데이터
      - Ex) 학년, 만족도, 선호도

중심 경향성

- 평균
  - 데이터의 총합을 데이터 개수로 나눈 값
  - 데이터가 대칭적인 경우에 사용된다
- 중앙값
  - 데이터를 작은 값부터 큰 값으로 나열했을 때 중앙에 위치한 값
  - 이상치가 많이 있는 경우에 사용
- 최빈값
  - 가장 빈번하게 나타나는 값
  - 범주형 데이터에서 주로 사용

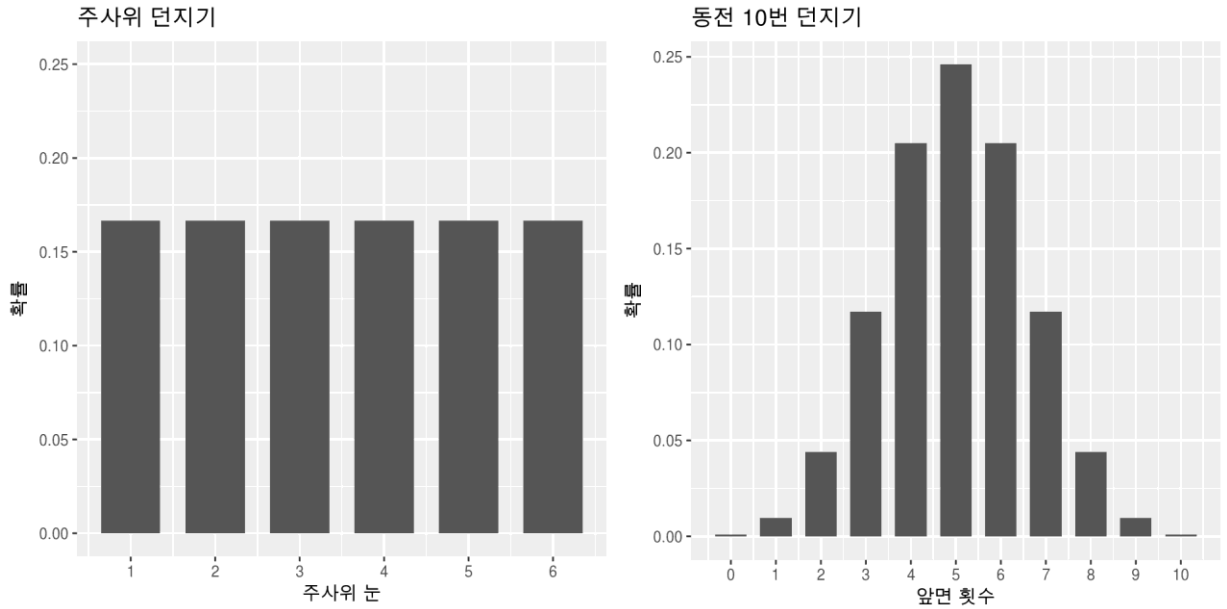


데이터 변이성

- 분산
    - 데이터가 얼마나 분산되어 있는지를 나타내는 값
    - 평균과의 거리의 제곱의 평균으로 계산
- 표준편차
    - 분산의 양의 제곱근
    - 분산과 함께 데이터의 변이성을 파악하는데 주로 사용

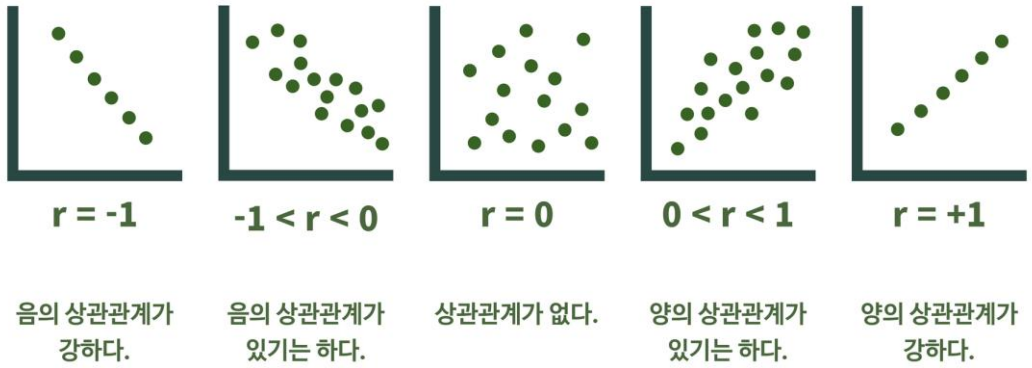
확률 & 확률 분포

- 확률
  - 어떤 사건이 발생할 가능성
  - 0과 1 사이의 값으로 표현
- 확률 분포
  - 확률 변수가 가질 수 있는 모든 값
  - 해당 값이 나타날 확률을 나타내는 함수
  - Ex)
    - 정규 분포, 이항 분포, 포아송 분포 등



상관계수 계산

- 상관계수 범위
  - -1 ~ 1까지의 값으로 나타낸다
  - 0에 가까울수록 두 변수 간의 관계가 약하다
  - 1에 가까울수록 강한 양의 상관관계가 있다는 것을 나타낸다



데이터의 척도

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

척도란?

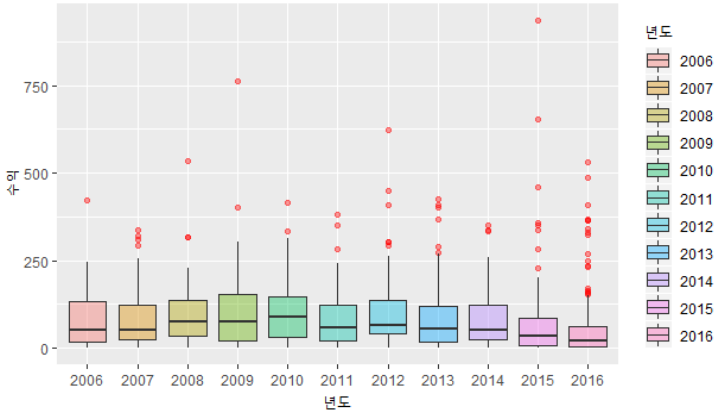
- 데이터를 측정하는 방법이나 척도를 적용하는 방식
    - 종류
        - 명목 척도
            - 데이터의 순서나 크기의 의미가 없는 척도
        - 순서 척도
            - 데이터의 순서는 의미가 있지만 크기의 차이는 비교할 수 없는 척도
        - 등간 척도
            - 데이터의 차이를 비교할 수 있는 척도
        - 비율 척도
            - 데이터의 비율이 의미를 가지는 척도

평균, 중앙값, 최빈값

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

평균

- 데이터의 총합을 데이터의 개수로 나눈 값
  - 평균 = (데이터1 + 데이터2 + … + 데이터n) / n
- 데이터가 이상치를 포함하고 있을 경우 평균이 대푯값으로 적절하지 않을 수 있다
  - 이상치
    - 극단적인 값, 잘못 입력된 값
- [이상치가 포함되어 있을 경우 중앙값이나 최빈값을 사용해 대푯값을 구하는 것이 적절]

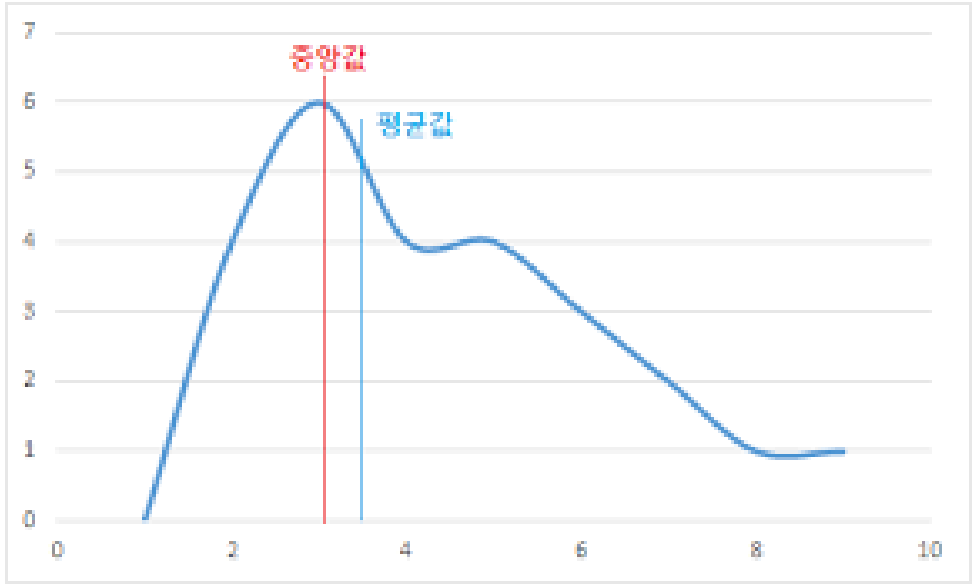


데이터를 대표하는 지표

중앙값

- 데이터의 중앙에 위치한 값
  - 중앙값 = (n + 1) / 2번째 값

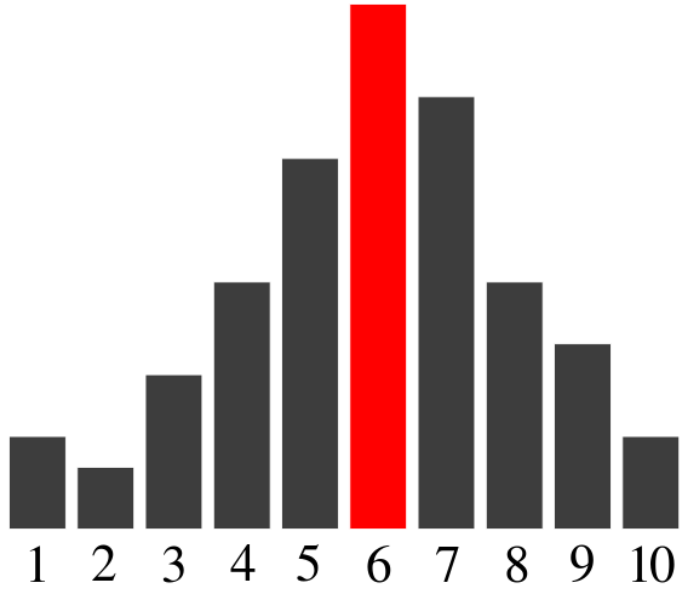
- Ex) 5명의 학생의 시험 점수가 각각 70, 80, 85, 90, 95점이다
  - 이들의 중앙값은?
  - 답 : 85



- 데이터가 홀수 개인 경우 중앙값이 정확하게 존재한다
- 데이터가 짝수 개인 경우 중앙값을 구하기 위해 가운데 2개의 값을 평균을 구한다

최빈값

- 데이터에서 가장 자주 나타는 값
  - Ex)
    - 1,2,2,3,3,3,4,6,6,6,6,6의 경우 최빈값 : 6



분산, 표준편차

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

분산이란?

- 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 지표
  - 식)
    - ((데이터1 - 평균)^2 + (데이터2 - 평균)^2 + … + (데이터n - 평균)^2) / n
    - ((데이터1 - 평균)^2 : 해당 데이터가 평균에서 얼마나 떨어져 있는지를 나타내는 차이값의 제곱

- 예시)
  - 5명이 학생의 시험점수가 각각 70, 80, 85, 90, 95이라면 분산은 다음과 같이 계산된다
  - 평균 = (70 + 80 + 85 + 90 + 95) / 5 = 84
  - 분산 = ((70-84)^2 + (80-84)^2 + (85-84)^2 + (90-84)^2 + (95-84)^2) / 5 = 74

표준편차란?

- 분산의 양의 제곱근
    - 데이터가 얼마나 퍼져 있는지를 나타내는 지표

- 수식 : 표준편차 = sqrt(((데이터1 - 평균)^2 + (데이터2 - 평균)^2 + ... + (데이터n - 평균)^2) / n)

분산과 마찬가지로 표준편차 또한 데이터가 평균에서 얼마나 떨어져있는지를 나타내는 것이다

다만, 원래 데이터의 단위와 동일하다 = 표준편차 값은 분산 값보다 직관적인 데이터 분포를 나타낼 수 있다

- 예시)
    - 5명이 학생의 시험점수가 각각 70, 80, 85, 90, 95이라면 표준편차는 다음과 같이 계산된다
    - 평균 = (70 + 80 + 85 + 90 + 95) / 5 = 84
    - 표준편차 = sqrt(((70-84)^2 + (80-84)^2 + (85-84)^2 + (90-84)^2 + (95-84)^2) / 5) ≈ 8.6

확률, 확률분포, 정규분포

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

확률이란?

- 어떤 사건이 일어날 가능성을 나타내는 지표
  - 확률은 0 ~ 1사이의 값으로 표현
  - 수식
    - P(A) = (A가 일어날 경우의 수) / (전체 경우의 수)

- 예시)
  - 주사위를 던졌을 때 1이 나올 확률
    - P(1) = (1) / (6) = 1/6

확률분포란?

- 어떤 사건이 일어날 확률을 나타내는 함수
  - 이산형
    - 확률변수의 값이 이산적인 값을 가질 때 사용
    - Ex) 동전 던지기를 했을 때 앞면이 나오는 횟수를 확률변수로 정의
  - 연속형
    - 확률변수의 값이 연속적인 값을 가질 때 사용
    - Ex) 학생들의 키를 확률변수로 정의
    - 정규분포, 지수분포 등

- 예시)
  - 주사위를 던졌을 때 1과 2가 나올 확률을 계산하고 싶다면?

정규분포란?

- 평균과 표준편차에 의해 결정
    - 평균값 : 분포의 중심
    - 표준편차 : 분포의 넓이와 산포도
    - 수식
        - f(x) = (1 / (σ * sqrt(2π))) * exp(-((x-μ)^2) / (2σ^2))
        - μ : 평균값
        - Σ :표준편차

- 정규분포의 그래프는 평균값을 중심으로 좌우대칭인 종 모양을 갖는다
    - 중심극한정리에 따라 매우 많은 데이터가 해당 분포를 따르게 된다
    - 데이터 분석에서는 대부분의 경우 정규분포를 가정하고 분석 수행

최종 정리

Department of Big Data in Software Engineering, Hallym University
Tae Hoon-Her

통계학 : 데이터를 수집, 정리, 해석하는 학문

몸무게와 같은 양적 데이터는 평균, 분산, 표준편차를 이용해서 분석할 수 있다

평균은 데이터의 중심을 나타내며 분산은 데이터가 중심에서 얼마나 멀리 흩어져 있는지를 나타낸다

- Ex) 5, 7, 8, 9, 10이라는 데이터가 있다고 가정
  - 평균 : (5+7+8+9+10)/5 = 7.8
  - 분산 : ((5-7.8)^2 + (7-7.8)^2 + (8-7.8)^2 + (9-7.8)^2 + (10-7.8)^2)/5 = 3.36

동전 던지기를 할 때 앞면이 나올 확률 : ½

정규분포는 대부분의 자연적 현상, 데이터 분포를 나타내는데 사용(평균과 표준편차에 의해 결정)

과제

학생들의 국어 성적 평균이 80점, 표준편차가 10점일 때, 100명의 학생 중 20명의 성적이 90점 이상일 확률은 얼마인가요?

[풀이과정을 함께 작성해주세요]

제출 기한 : 2023/05/18(목)

제출 방법 : 자필 or 태블릿에 작성 후 스캔하여 제출