통계 1주차 정리

표본 : 현재 가지고 있는 데이터
모집단: 아직 가지고 있지 않은 모르는 데이터

-> 통계는 표본이라는 일부 데이터를 이용해서 모집단이라는 전체 데이터를 분석

표본추출(sampling) : 모집단에서 표본을 얻는 것을 말함
단순무작위표본추출(simple random sampling)
ex) 무작위
계통추출(systematic sampling)
ex) 모집단에 순서를 매기고, 일정 간격으로 표본추출
층화추출(stratified sampling)
ex) 여러 층 (10대, 20대, 30대)를 나누고 층에서 무작위 추출
군집추출(cluster sampling)
ex) 지역별로 학교 나누고 , 몇 개의 학교만 선택하여 조사

-> 하지만 최근 트렌드는 가지고 있는 데이터에서 표본을 뽑는 행동은 하지 않음, 될 수 있는 한 많은 데이터를 사용

확률 :어떤 데이터를 얻을 수 있는 확률
확률분포: 확률분포 확률변수와 그 값이 나올 수 있는 확률을 대응시켜 표시

	수치형(numerical)	연속형, 이산형
변수		
	범주형(categorical)	순위(서열)척도, 명목척도

수치형 : 정량적인 수치 ex) 개수 , 나이
이산형 : 키 , 몸무게

서열척도 : 순위 , 만족
명목척도 : 혈액형 , 키를 그룹으로 설정 명목척도로 만들 수 있음

why 변수에 대해서 잘 알아야 하냐 -> 분석 방향이 달라짐

계급값 : 범위 내의 최댓값과 최솟값의 중간값
도수 : 데이터가 나타난 횟수 , 빈도
도수분포
상대도수 : 전체를 1로 두었을 때 도수가 차지하는 비율
-> 엔트로피 개념 / 분류 classification에서 중요한 개념

히스토그램 : 도수분포를 도표로 나타낸 것

통계량 : 데이터 aggregate 한 값
- 평균값
- 기댓값 : 확률 x 값 , 데이터를 손에 넣지 못했다고 해도 확률분포를 알고 있다면 기댓값 계산 가능
- 분산 : 데이터가 평균값과 얼마나 떨어져 있냐 , 데이터가 모여 있으면 분산이 작아짐 멀리 떨어져 있으면 분산은 커짐

모집단 분포 추정
모집단의 분포는 알 수 없음 , 모집단 분포는 추정하는 것이 전제 -> 추측통계
우리는 항상 표본으로 모집단을 가정해야 함
이것은 모집단의 분포와 표본의 분포가 같다는 가정이 들어감

왜 정규분포(가우스 분포)를 사용하는 것인가?
1. 실험이나 관찰을 통해 수집된 데이터의 확률분포는 대부분 좌우 대칭이며 종형 분포를 보이고 있어 정규분포를 따르기 때문
2. 정규분포를 하지 않는 변수들의 경우에는 변환(제곱근, 세제곱근, 로그 등)을 통해 정규분포에 근사하도록 유도가 가능
3. 정규분포는 평균과 표준편차만 주어지면 정의되고 수학적으로도 편하게 계산되며, 여러 다른 분포들과 긴밀한 관계를 맺고 있음

확률 질량 함수 vs 확률 밀도 함수
확률 질량 함수: 이산확률 변수 x가 어떤 값 x를 취할 때의 확률을 대응하는 함수
확률 밀도 함수 :연속확률 변수 x의 확률 밀도를 나타내는 함수

확률 질량 밀도 함수의 경우 데이터의 성질에 맞게 각 확률을 구하기 위한 함수

모수 : 모집단의 특징을 나타내는 수치
모수는 평균, 분산이 있음

모집단분포를 정규분포라고 가정한 후 -> 중심극한정리
 정규분포의 모수(평균과 분산)을 구할 수 있으면 모집단의 분포를 추정할 수 있음
-> 표본통계를 모집단분포의 모수라고 생각

추정오차 : 표본의 통계량을 모집단 분포의 모수라고 생각하기 때문에 실제 표본의 통계량과 모수에는 차이가 있음. 추정된 모수에는 추정오차가 존재함
이를 위해 구간추정 방법 , 통계적 가설 검정 사용

확률 밀도 함수 : 해당하는 구간의 적분을 통해 확률 값을 구함

$$P(a \le X \le b) = \int_a^b f(x)dx$$

확률 질량 함수 :확률값의 경우 해당하는 확률 값 만 더하면 됨

$$P(1 \le x \le 3) = \sum_{i=1}^{3} f(x_i)$$

어떤 확률변수 X가 평균(기댓값) u, 분산 의 정규 분포를 따른다.

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

표시