|  |  |
| --- | --- |
| 교육 제목 | 통계학에 따른 자료의 요약 |
| 교육 일시 | 2021.10.27 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 오전 | ●자료의 종류는 연속형(회귀)와 범주형(이산형)으로 나뉜다  ◎연속형자료  등간척도 : 온도가 대표적이며 절대적인 無값이 존재하지 않는다  비율척도 : 키 몸무게 혈압이 대표적이며 無값이 존재한다  ◎범주형자료  명목척도 : 속성간 격차가 없고 고유의 뜻을 가진다 ex) 남/녀 , 국어/수학  순서척도 : 속성간의 격차가 있어 순위가 발생함 ex) 상/중/하 , 수/우/미/양/가  ●자료의 축약  ◎모집단 : 이해하고자 하는 집단 전체  ◎표본집단 : 모집단에서 추출된 집단, 모집단의 좋은 추정치가 된다  빅데이터라는건 표본을 추출한다 해도 모집단의 규모가 된다  ◎도수분포표 : 각 범주에 속하는 계급값과 도수를 표로 나타냄  계급값은 그 범주의 평균값을 뜻하고, 도수는 그 범주의 개수를 뜻한다  ◎히스토그램 : 그래프와 확률밀도함수(물결형그래프, PDF로도 쓴다)가 혼합된 데이터형  ●그래프 요약의 장단점  장점 : 한눈에 알아볼 수 있다  단점 : 작성자에 따라 주작가능  ●자료의 요약  ◎표본평균(표준편차) : 중심이 있는 축도에서 가장 많이 사용하는 방법이며 자료의 무게중심을 나타낸다  확률밀도의 무게가 같아지며 지료의 이상치에 영향을 많이 받는다  ◎중앙값 : 크기순으로 나열할 때 중앙에 위치한 값  자료의 이상치에 영향을 적게 받는다  이상치는 위 두 값에 많은 영향을 준다  ◎최빈값 : 가장 자주 나오는 값 |
| 오후 | ●퍼짐의 정도 : 자료가 중심으로부터 얼마나 퍼져있는가?  ◎범위는 최댓값에서 최솟값을 뺀다  ◎사분위수(범위는 중앙값과 같다)는 삼사분위에서 일사분위를 뺀다  이상치의 한계식 : Q3 + 1.5 x IQR  하방의 식은 +가 -로 바뀐다  ●변수간 자료요약  ◎상관분석 : 두 연속형 변수간 연관관계를 분석해낸다  ◎피어슨의 상관계수  두 연속변수가 정규분포를 따를 때 좋다  r값이 -1에 가까울수록 음의 상관성  +1에 가까울수록 양의 상관성을 가진다  또한 절댓값이 1에 가까울수록, 일련의 패턴이 있을수록 상관성이 높다고 따진다  ◎스피어만 상관계수  두 연속변수가 정규분포를 따르지 않을 때 사용 |