|  |  |
| --- | --- |
| 교육 제목 | **통계와 확률** |
| 교육 일시 | 2021년 9월 15일 |
| 교육 장소 | 영우글로벌러닝 |
| **교육 내용** | |
| 오전 | 통계는 값이 정해진 범주형 자료와 값이 정해지지 않은 연속형 자료가 있는데 여기서도 세부적으로 나눠지는데 범주형 자료는 따로 순위나 순서가 없는 명목척도와 순서나 순위가 있는 순서척도로 나뉘고 연속형 자료는 절대값0이 존재하는 비율척도와 절대값0이 존재하지 않으며 곱셈법칙 적용이 되지 않는 등간척도로 나뉜다. 통계를 위한 자료는 조사하고자 하는 목표집단 전체는 모집단, 모집단에서 추출된 집단은 표본집단으로 나뉜다. 자료를 보기 쉽게 정리할 때 자료종류에 따라 표시방식이 나뉘는데 범주형 자료는 분할표로 간편하게 표현할 수 있지만 이건 상대적으로 보기가 불편해 대부분 막대로 표시되는 막대그래프나 원 등을 나누는 파이그래프를 쓴다. 하지만 연속형 자료는 연속되는 특성상 나눠 표현할 수가 힘들에 돗수분포표(범주를 정해 관측값의 개수를 적어서 표시한다.)나 최솟값과 중간값(Q2), 최댓값과 25%값인 Q1, 75%값인 Q3까지 표시한 상자 수염 그림, 모양이 바이올린과 닮은 모양이 있어서 이름붙은 바이올린 그림 등으로 표시한다. 앞에서 보기 쉬워서 그래프를 쓴다고 했지만 한 눈에 보기 쉽게 만든다는 것 때문에 모양이 만드는 사람의 주관에 따라 달라질 수 있어서 마음만 먹으면 악의적으로 왜곡할 수 있다는 단점도 있다.  중심위치를 표시하는 방법들이 있는데 자료의 무게중심을 표시하는 표본평균(평소 많이 듣는 표준과 비슷하거나 같다.)이 이해가 쉬워 가장 많이 쓰이지만 이상치(자료의 대체적인 흐름에서 많이 벗어난 값)의 영향을 많이 받는다. 그래서 이상값에 영향을 적게 받는 중앙값(자료를 크기 순으로 나눴을 때 중앙에 있는 값)으로 보완을 한다. 또 관측값 중 가장 자주 나온 값인 최빈값도 쓰인다.  또 분산과 표준편차는 자료가 중심에서 얼마나 퍼져있는지를 표시하는데 표준편차는 분산의 제곱근값이다. 범위는 우리가 흔히 아는 최댓값-최솟값이고, 사분위수범위는 위에 설명한 75%인 Q3에서 25%인 Q1을 뺀 값이다. |
| 오후 | 상관분석은 두 연속형 변수간에 연관된 정도를 분석하는 통계인데 이것이 인과관계를 나타내는 것은 아니므로 주의해야 하고 상관계수가 이 분석에서 선형적 관계의 강도를 나타낸다. 여기서 모집단 상관계수는 p, 표본집단의 상대계수는 r로 표시한다. 그리고 그림으로 표시할 때는 산점도를 사용한다. 그 중 피어슨 상관계수는 두 변수가 정규분포를 따를 때 사용하는데 r값이 양수면 양의 상관관계, 음수면 음의 상관관계이다. r은 -1에서 1까지 있는데 절대값이 1, 즉 1이나 -1에 가까울 수록 상관성이 높다. 또 스피어만 상관계수는 두 변수 사이의 상관관계를 비모수적인 방식으로 계산, 표시하는 방식인다 두 변수가 정규분포가 아닐 때 사용하고 비선형적인 자료도 연관성이 있는지 알 수 있고 r의 범위는 피어슨의 상관계수처럼 -1에서 1까지 있고 절대값이 1에 가까울 수록 상관성이 높다.(원하면 순서형 자료에도 쓸 수 있다.)  확률은 어떤 일이 실제로 일어날 것인가를 보여주는 건데 무한히 반복될 때의 값이라 적은 개수는 확률에서 벗어날 가능성이 높다. 두가지 종류의 값이 쓰일 때 조건부 확률과 독립인 경우일 때 두 확률이 동시에 일어날 확률이 두 확률을 곱한 것과 같다는 게 중요한데 실제 자료는 양이 많아 직관적으로 눈에 들어오지 않아 이것으로 독립인지를 구분해야 될 수 있기 때문이다. |