2021.09.15.

통계의 기초

통계학과 데이터사이언스와의 차이: 기존의 통계학은 데이터가 정형화 되어있었다, 예를 들어 나이, 지역, 나라, 인종, 연봉, etc... 하지만 요즘의 데이터사이언스는 데이터가 비정형인 것들이 많다, 예를 들어 사진, 웹에서 긁어온 다량의 데이터...etc 이러한 것들은 기존의 통계학으로는 다룰 수가 없고 컴퓨터프로그래밍이 가능/불가능의 차이도 있다.

데이터에는 크게 정형과 비정형 자료가 있다. 정형 데이터에는 몇 가지 종류가 존재한다. 연속형 자료, 범주형/이산형 자료가 있다. 연속형 자료는 등간척도/interval과 비율척도/ratio가 있다. 등간척도는 절대값 0이 없는 자료로 대표적인 예로 온도가 있는데 이때의 0도는 수로 표현했을 때 0이지만 물리적으로는 영하와 영상의 온도의 중간일 뿐이다. 비율척도는 절대값0이 존재하는 자료로 예로 무게와 압력 등이 있는데 이때의 0은 물리적으로 질량/무게가 없거나 가해지는 힘이 없는 경우다. 명목척도/nominal과 순서척도/ordinal이 있다. 명목척도는 속성을 분류하는 것으로 대표적인 것으로 교과목으로 국어/영어/수학이 있다. 순서척도는 속성 간에 차이가 존재하는 것으로 대표적으로 국어-수 영어-미 수학-양 이 있다. 비정형 데이터는 말 그대로 정해진 것이 없는 자료인데 예를 들어 사람의 모습, 자연의 풍경, 냄새, 이러한 것들이 있다.

어떠한 것을 분석하고 이해하려는 과정 중에는 자료의 수집이 필수적이고 그러한 자료를 모을 때 모든 자료를 모으고 나타내는 것이 이상적이겠으나 물리적으로 힘든 경우가 많고 그를 위해 데이터를 축약/요약한다. 데이터를 이해/분석하기 전에 어느 것의/에서 데이터를 추출할지를 정해야 하는데 그 때 자료를 추출을 할 집단을 모집단이라고 하고 예를 들어 우리나라 남자의 평균 키가 있다. 자료를 추출할 때 모든 데이터를 얻으면 좋겠으나 현실적으로 불가능한 경우가 많고 그 경우에는 모집단에서도 일부 집단을 정해서/뽑아서 데이터를 추출을 하는데 우리나라 남자의 평균키를 예로 들었을 때 10대 남성 50명 20대 50명...60대 50명 이러한 경우를 예로 들 수 있다. 주의 할 것으로는 데이터를 추출할 집단을 잘못 설정하거나 표본 집단을 여러 방면으로 고려하지 않으면 결과가 왜곡 될 수가 있다, 예를 들어 사람의 암을 치료하기 위한 연구를 할 때 표본 집단을 대한민국에서만 선정해서 데이터를 추출한다면 연구 결과를 다른 나라 사람들한테는 유전자나 여러 나라의 사람들의 다른 생활 방식 등의 여러 면으로 대한민국 사람들과는 다른 변수로 인해 적용하기 어려울 수 있다.

데이터에서 정보를 뽑아내고 그것의 의미를 이해하기 위해서는 데이터의 구조와 크기를 잘 파악하고 분석을 해야 하고 그것을 적절하게 요약을 할 수 있어야 한다. 데이터의 구조와 크기를 잘 파악한 다는 것은 어떠한 방법으로 데이터를 추출하는 가를 말하는 것인데, 예를 들면 대한민국의 사람들의 재산을 알기위해 우리나라의 모든 사람들을 대상으로 데이터를 추출할 텐데 이때 재산을 유동성을 기준으로 범주를 나눠 데이터를 추출할 수도 있고 모든 재산을 원화 화폐로 바꿔서 데이터를 산출할 수가 있고 그 데이터에서 나오는 정보를 잘 이해하려면 요약/표현도 적절하게 해야한다. 데이터를 표현 할 때 그래프를 통해 데이터를 요약할 수도 있고 수치를 통해 요약을 할 수가 있다. 범주형의 자료의 경우 요약 방법 중 대표적인 것은 돗수분포표/frequency table, 막대그래프/bar graph/bar plot, 파이차트/pie chart가 있다. 연속형의 경우에는 돗수분포표 (Frequency table), 히스토그램 (Histogram), 상자수염 그림 (Boxplot), 바이올린 그림 (Violin plot)을 대표적인 예로 들 수 있다. 그래프를 통해 정보를 표현 했을 때에는 직관적으로/한 눈에 이해하기가 편하다. 하지만 그 그래프를 작성한 주체가 기준이나 그래프의 모양, 수치의 표현을 통해 정보의 의미를 왜곡 시킬 수 있다는 주의할 점이 있다. 간단한 예를 들자면 미국을 표현하기 위해 미국의 생겨난 날부터 지금까지의 1인당 국민소득을 막대그래프로 나타내면 굉장한 성장세로 현재의 42000달러에 달하는 것을 보게 되어 미국을 굉장히 살기 좋고 국민 모두가 부자로 지낼 것만 같은 의견을 가지게 될 수가 있는데 이때 빈부의 격차나 세계적인 사건으로 인한 미국 경제의 성장과 같은 기준들과 변수가 배제되어 미국에 대한 생각 왜곡될 수가 있는 것이다. 수치를 통해 연속형 자료를 요약하는 방법들 중 중심위치를 측정하는 것들로는 대표적으로 표본평균 (Sample Mean), 중앙값 (Median) 그리고 최빈값 (Mode)이 있다. 표본평균은 중심을 나타내는 측도 중에서 가장 많이 사용되는 방법으로 자료의 무게 중심을 나타내는데 공식은 이고 자료의 이상치의 영향을 많이 받는다. 중앙값은 전체 관측값을 크기순으로 나열한 했을 때 중앙에 위치한 값을 나타내고 공식은 이고 데이터의 수가 홀수면(n+1)/2번째 관측값을 쓰고, 짝수이면 n/2번째와 (n-1)/2번째 관측값의 평균값을 쓰고 자료의 아상치에 영향을 적게 받는 다는 점이 있다. 최빈값은 관측값 중에 가장 자주 나온 값을 말한다. 자료에 이상치의 정도와 개수에 따라 표본평균과 중앙값을 적절히 써야 한다, 그렇지 않으면 왜곡된 결론이 나올 수 있다.

데이터를 통해 값을 관측하고 거기에서 의미를 이끌어 내기 위해서는 변수들과 데이터의 관계를 파악하는 것이 매우 중요하다. 이때 두 연속형 변수간에 선형적 연관관계가 있는지 분석하는 통계적 방법을 상관분석(correlation)이 있다. 하지만 상관분석은 두 변수간에 연관된 정도를 나타내는 것이고 인과관계와 같은 것은 아니다, 예를 들어 아이스크림의 판매량이 증가했더니 상어에 의해 사망한 사람들의 수가 증가하는, 즉, 양의 상관관계를 나타내는 통계가 있다고 가정하자, 그러면 아이스크림의 판매량과 상어에 의해 사망한 사람들의 수는 서로 비례하여 아이스크림이 많이 팔리면 사람들이 상어한테 죽는다는 말이 안 되는 관계가 나타난다, 하지만 사실은 여름이라 아이스크림 판매량이 증가한 것이고 여름이라 해안가에 상어가 자주 출몰하여 피서를 온 사람들이 상어에 의해 피해를 입은 수가 증가한 것으로 두 사건들은 서로 전혀 상관이 없는 독립적인 사건들이고 인과관계가 있는 관계라면 여름과 아이스크림 판매량, 그리고 여름과 상어의 출몰의 빈도가 되겠다.   
 위에서 상관관계가 언급되었는데 통계에서 상관관계는 두 변수가 있을 경우 계수로 값을 구할 수 있으며 양의 상관관계와 음의 상관관계를 판별할 수 있다. 두 연속형 변수가 정규분포를 따를 때에는 피어슨 상관계수(Pearson‘s r)를 이용하여 상관관계를 나타낼 수 있다. r값이 –면 음의 상관관계이고 +이면 양의 상관관계인데 둘 다 1에 가까울수록 높은 상관성을 뜻하고, 낮을수록 낮은 상관성을 뜻한다. 공식은 이다. 두 연속형 변수가 정규분포를 따르지 않는 경우에는 비모수적 방법으로 상관관계를 나타내는 스피어만 상관계수(Spearman‘s r)을 사용할 수 있다. 이 계수는 순서형 자료에서도 적용이 가능하고 비선형적 연관성도 판단이 가는하고 공식은 이고 상관성은 피어슨 상관계수와 동일하게 –1부터1까지 있으며 절대값 1에 가까울수록 높은 상관성을 뜻한다.

어떠한 현상의 데이터를 다루면서 그 현상이 발생할 확률을 구하는 것은 필연적이다. 확률은 동일한 실험을 무한히 반복했을 때 나타나는 사건의 상대돗수의 비로 100분율을 이용해 나타낼 수 있으며 0보다 작을수 없고 1(100%0)를 넘을 수 없다. 확률을 계산하는 방법으로는 여사건의 법칙P(Ac)=1−P(A), 합사건의 법칙P(A∪B)=P(A)+P(B)−P(A∩B), 합사건의 여사건P((A∪B)c)=P(Ac∩Bc), 조건부확률 과 두 사건이 독립인 경우의 P(A∩B)=P(A)P(B)으로 계산을 할 수 있다.