1. 인과관계와 상관관계?

인과관계	상관관계
• A → B • A를 하면 '항상', '언제나' B가 발생하는 명확한 관계성을 지니고 있다. • 원인과 결과의 순서가 바뀔 수 없는 비대칭적 관계를 가지고 있다.	• A & B • A와 B의 연결성으로, 두 변수들이 얼마나 상호 의존적인지를 의미한다. • 연결성을 중심으로 보기에 요소 간 순서가 무의미하므로 대칭적 관계

▸ 인관관계와 상관관계가 동일할까? NO

나비효과를 생각해보면 된다. 나비효과란 나비의 날갯짓이 지구 반대편에선 태풍이 될 수도 있다는 의미이다. 하지만 실제로 나비의 날갯짓으로 태풍이 일어나는가? 그렇지 않다. 기류와 날씨 변화, 해류 등 '수많은 요인'이 '조합'되어 태풍이 일어나기에 나비의 날갯짓이 시작점이 될 순 있지만, 필수불가결한 원인은 아니다.

▸ 데이터 기반 의사결정에서의 인과관계와 상관관계

최근 기업에서는 '데이터 기반의 의사결정'을 강조하고 있다. 데이터 기반의사결정이란 숫자 간에 맺어지는 관계를 토대로 결론을 짓는 행위를 말한다. 하지만 이때 유의해야 할 점은 **'상관관계가 인과관계를 나타내지 않는다'**는 점이다. 하나의 결과에는 수많은 원인 요소를 이루어졌을 수 있기에 하나의 요소가 '절대적' 인과성을 지니지 않음을 알아야 한다. 빅데이터 시대를 맞아 이전과 달리 데이터를 토대로 인과성이 아닌 상관관계성을 파악하려는 흐름이 주류이다. 이때 이러한 상관성을 숫자로 표현한 것이 바로 상관계수이다.

하지만 상관관계를 분석하는 과정에서도 유의해야 할 점이 있다. 상관관계가 많이 사용되고 있는 만큼 오용되고 있는 경우도 적지 않다. 이를테면 소아마비와 아이스크림 섭취와는 아무런 연관성이 없었음에도 불구하고 이 둘의 결과치를 중심으로 상관관계로 1940년 보건 발표가 있다는 점이 이를 증명한다. 이러한 오류가 발생하는 이유는 바로 아무런 연관성이 없는 외생변수를 관계성이 있다고 보아 둘 사이의 연관성을 잇고자 하는 무분별한 의미 부여와 패턴 찾기로 비롯된다. 그러므로 데이터를 분석하는 과정에서 내생 변수와 외생 변수를 정확히 이해하고 변수의 추이에 영향을 미칠 수 있는 '실질적' 요인이 무엇이 있는지 꼼꼼하게 살펴볼 수 있어야 한다.

2. 숫자의 불확실성

인과관계와 상관관계에 대한 구분이 이루어졌다면, 숫자의 불확실성에 대해서도 이해해야 한다. 숫자는 **'정량적인 비교'**이다. 우리가 1, 30, 2.5 등 표기된 숫자를 읽는 방식은 사실상 '절대적' 의미를 지닌다고 할 수 없다. 그저 수많은 사람의 공통된 전제 하에 만들어진 **'기준'**에 불과하다. 이를테면 1개가 든 바나나 봉지와 3개가 든 바나나 봉지를 보고 우리는 모두 '바나나 한 봉지'라고 표현한다. 봉지 안에 든 바나나 개수보다 봉지의 단위를 우선으로 읽게 된다. 또 친구가 옆에서 말하는 '한 입만'은 내 기준의 '한 입'과 다를 수 있지 않을까? 이를 데이터로 끌어본다면 다양한 요소가 결합되어 결과물을 산출하는 데이터의 특성을 이해하고 수치를 바라볼 수 있어야 한다. 데이터 분석가는 숫자 간의 단순한 차이가 발생하더라도, 그 차이가 '통계적으로 의미가 있는지', 고려해야 할 전제 조건은 없는지, 그 차이가 정말 여러 측면에서의 차이가 맞는 것인지를 여러 도구(예: t-검정, z-검정)를 사용하여 고민할 수 있어야 한다.

3. 모수와 표본

모수	표본
• 모집단의 수치적 요약값 (즉, 단순히 모집단의 수가 아니라, 모집단의 통계값 예: 모평균, 모표준편차 등) • 모수가 중요한 이유? 모수의 값을 근거로 모집단의 형태를 추정할 수 있기 때문 • 현실에서 모수를 다룰 수 있을까? NO why? 전체 데이터를 다 사용한다 한들, 그 데이터가 다양한 케이스를 모두 대표 불가 ※ 모집단 : 분석하려고 하는 대상 전체 집단	• 모집단에서 추출한 일부 데이터로, 모집단을 대표할 수 있는 일부 정보를 담고 있다. • 표본은 모집단을 분석하거나 모집단에 대한 추론을 하기 위해 사용한다. • 큰 수의 법칙 **:** 표본의 크기가 충분히 크다면 그때의 표본평균은 모평균에 충분히 가까워진다**.**

4. 확률과 분포

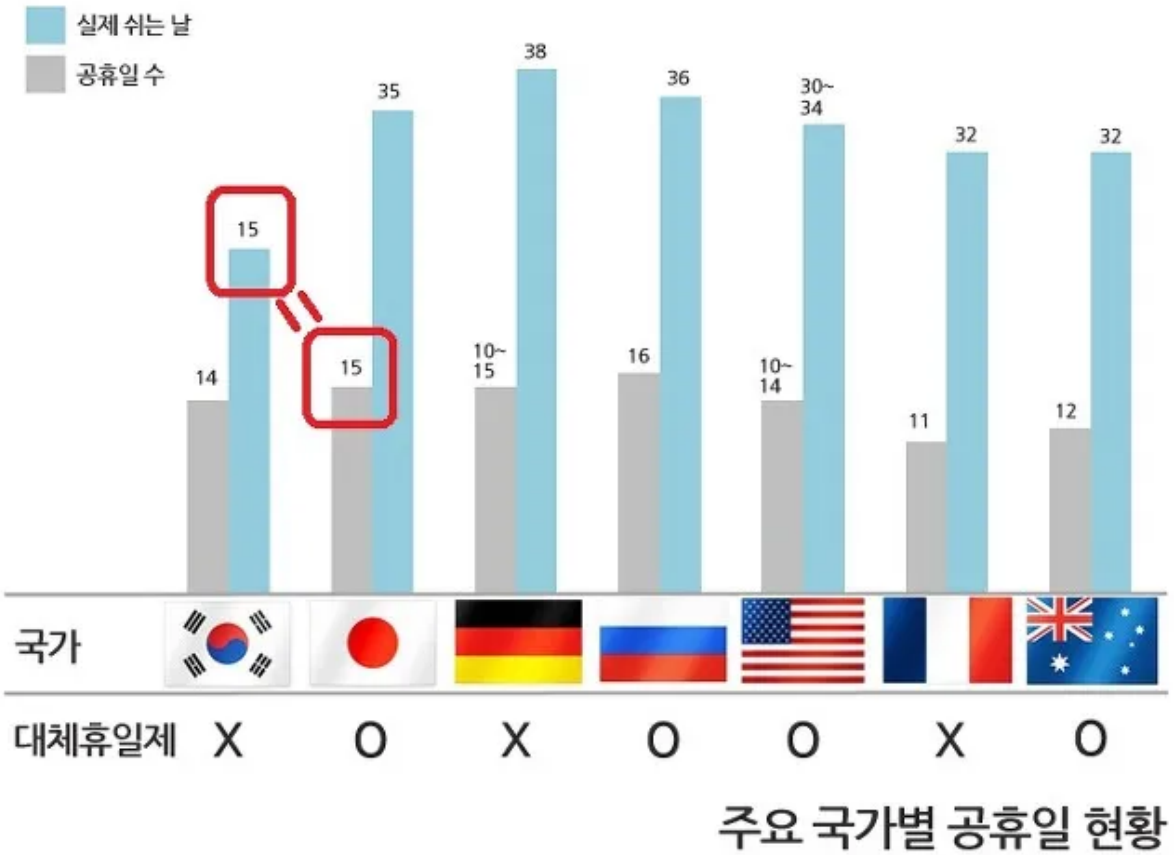
확률	분포
• 일정한 조건 아래에서 어떤 사건이나 사상이 일어날 가능성의 정도로 시행 결과값의 평균을 **0**과 **1** 사이로 나타낸 것을 의미한다. • 과거의 데이터를 토대로 앞으로 일어날 일을 예측한다.	• 확률이 어떤 모습으로 퍼져 있는지를 나타내는 것을 의미한다. • 확률분포를 의미하며, 확률 변수가 특정한 값을 가질 확률을 나타내는 함수이다.
※ 확률의 개념에서 유의해야 할 점 : 확률은 지나간 사건의 결과를 보장해주지 않는다. 즉, 확률이 10%라고 하여 결과값이 언제나 10%에 수렴하는 것이 아니다. '대체로 그럴 가능성이 높다'라는 의미에 지나지 않는다. ▸ 앞서 실행된 결과 데이터가 없는 상태에서 확률을 구하고 싶을 땐 어떻게 해야 할까? **"**테스트 진행**"** 사건을 임의로 일으켜 데이터 집합을 만든다 (예: A/B 테스트) (예) 홈페이지의 새 배너를 보고 구매행동을 할 확률 → A : 새 배너를 적용한 집단 vs. B : 기존 배너 적용 집단	

5. 실험을 통한 의사결정

많은 기업은 '실험'을 통해 '데이터'를 수집하고 '의사결정'을 한다. 예를 들어 온라인 서비스에서 실험을 한다면 버튼의 모양을 바꾼다거나 사용자의 UX를 변경하며 A/B테스트를 이용하는것을 예시로 들 수 있다. (이때 유의해야 할 점은 테스트를 진행할 때 변수가 적을수록 좋다**! why?** 변동사항이 많다면**,** 어떠한 이유로 변화가 일어났는지 예측하기 어렵기 때문이다) 그렇다면 이러한 테스트를 하는 이유는 무엇인가**?** 바로 **'**기존에 없던 기록**'**을 얻기 위함이다. 기업에서 새로운 변화를 반영할 때 가능한 데이터를 확보하여 불확실성을 줄이고 싶어하기 때문이다. 하지만 그렇다하여 실험이 모든 것을 해결해주는 것은 아니다. 실험 대상이 된 데이터는 다른 데이터와 마찬가지로 고객의 사용 내역이 기록된 데이터이지만, 실험 내용이 섞여 있으므로 그대로 사용할 수 없다.

6. 그래프 읽기

숫자로 표현된 상황이나 현상에서 빠르게 얻기 힘든 통찰을 훨씬 쉽게 얻는 데 잘 만들어진 그래프만큼 좋은 도구도 없다. 하지만 그래프 또한 그 이면의 숫자와 만든 이의 의도가 담겨져 있기에 '객관적 현상 이해'로만 바라보아서는 안 된다. 아래의 이미지처럼 왜곡된 그래프가 많이 발생하기도 한다.



주요 국가별 공휴일 현황

그러므로 데이터분석을 통한 시각화 과정에서 그래프는 데이터의 결과물을 예쁘게 만든다는 점도 있지만, 근본적인 목적은 '데이터를 직관적으로 이해할 수 있게 한다'는 것임을 염두해두어야 한다.

7. 추세선 그리기

추세선은 모든 데이터에 사용 가능한가? NO!
추세선이 사용 가능한 경우는 언제일까? (a) 데이터의 X축이 '일정한 시간 단위'일 때 (예: 과목별 점수는 불가능, 1~12월은 가능) (b) 추세선의 정확도 (이때 참고할 수 있는 것 중 하나가 R-제곱값) ※ R-제곱값: 0 ~ 1 사이의 값으로 추세선과 실젯값이 얼마나 비슷한지를 나타낸다 (이때, 1에가까울수록 추세선과 실젯값이 비슷하다고 할 수 있다. 또한, R-제곱이 0.1도 안되는 추세선은 신뢰도가 많이 낮을 것이라 판단할 수 있다.) ▸ 추세선을 알맞게 활용하기 추세선은 말 그대로 '추세'를 보여주는 선이다. 이러한 추세를 정확하게 그래프로 나타내야 하며, 일정하지 않은 시간 단위의 데이터를 사용하면 추세선은 왜곡된 형태로 나타나게 된다. (예: 1월 상반기, 2월 하반기 등) 즉, 추세선을 그랠 때는 추세선의 '정확도'를 신경써야 한다.

8. 시계열 데이터

- 시간에 따른 변화를 데이터로 나타낸 것
- 주로 '추세', '주기', '계절성'으로 구분하여 분석한다.
    - 추세 : 장기적으로 늘어나거나 줄어드는 형태
    - 주기 : 고정된 시간 단위로 유사한 변동 형태가 나타나는 경우
      (예: '24시간'이라는 일정 시간을 기준으로 평일의 지하철 이용 행태 분석
      → 출퇴근 시간의 혼잡도가 가장 높을 것 = 주기)
      ※ 주기의 변화가 발생한다는 점 → 향후 추세(장기성)가 변화할 수 있다는 의미 시사
    - 계절성 : 주기적으로 반복되는 때에 어떤 사건이 발생하는 것을 의미
      (예: 매년 빼빼로데이에 빼빼로 판매량이 증가하는 것 → 계절성의 예시)

▸ 주기 vs. 계절성의 비교
주기와 계절성은 둘 다 '반복'된다는 점에서 공통되어 보일 수 있다.
하지만, 주기는 '형태'의 반복이고, 계절성은 '빈도'의 반복에 가깝다.
(예) 지하철 승객 수는 24시간이라는 주기 안에서 동일한 '형태'가 반복되는 것을 의미하고, 빼빼로 판매량의 '증가'라는 사건의 반복은 11월 11일이라는 계절성을 반영하고 있다.

9. 별점의 함정

데이터의 유의점 : 데이터는 무슨 일(what happen?)이 일어났는지를 알려줄 순 있어도, 그 일이 '어떻게(how?)' 일어났는지 혹은 어떤 감정으로 일어났는지는 알려주지 않는다. → 설문조사 결과 사용의 한계점
그렇다면 어떻게 고객의 만족도를 확인할 수 있을까? 사용자가 직접 데이터를 입력하는 설문조사의 방식 대신 고객의 실제 행동 데이터와 같은 '프로그램이 남기는 데이터'를 보고 추정하면 된다. (예: 고객의 재구매율)

10. 인구통계학 정보의 효용성

▸ 페르소나 방법론 : 서비스나 UX 기획에 사용되는 기법으로, 가상의 고객을 구체적으로 정의하고, 이 고객이 서비스를 어떤 필요에 의해서 어떤 식으로 사용할지 구체적으로 그려보는 방식이다. 이를 통해 기획자는 본인의 의지 개입 여지를 줄이고 고객의 의도를 더욱 잘 이해할 수 있다는 장점이 있어 유용하게 사용된다. 페르소나 방법론은 고객의 나이, 성별, 지역 등과 같은 인구통계학적 정보만 다루기보다는, 사람의 안에 있는 수많은 다양성를 기반한 행동 데이터를 중점으로 다룬다.

11. 조건부 확률

- 주어진 조건의 발생 여부에 따라 확률이 달라지는 경우

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

12. 범위 제한을 통한 정확도 향상

데이터 활용의 근간은 바로 '논리'에서 시작된다. 데이터 분석은 '의사 결정의 근거'를 만드는 것이고, 결국 그 근거를 토대로 '의사 결정자'들을 설득해야 하기 때문이다. 그러므로 명확한 정의와 범위의 제한을 통해 '논리'를 구축할 수 있어야 한다.

13. 평균

일반적으로 '평균'은 전체 집합의 값을 더한 후, 그 잡합의 원소 개수로 나누는 '산술평균'을 의미한다.

▸ 평균의 함정 : 평균은 모든 것을 보여주지 않으므로 상황에 따라서는 평균이 아닌 다른 대푯값도 함께 살펴볼 필요가 있다. 대푯값인 평균은 진실을 잘 '요약'해주긴 하지만, 모든 진실을 완벽하게 반영하는 것은 아니다. 그렇기에 데이터 분석의 과정에서 우리는 그 분포가 어떻게 생겼는지도 고민할 수 있어야 한다.

※ 대푯값 : 주어진 집단을 요약해서 나타낼 수 있는 값으로 평균, 최빈값, 최댓값 등이 있다.

14. 데이터 문해력

- 일반적인 문해력처럼, 데이터를 사용해서 '읽고 쓰고 말하고 듣는' 능력을 의미한다.
- HOW TO? 데이터를 올바르게 바라보는 방법
    - 데이터의 출처와 목록을 파악하기
    - 데이터에서 누락된 부분이 없는가?
    - 논리에 허점이 없는지 확인하기

15. 통계 용어

- DAU(Daily Active Users) : 하루 동안 해당 서비스를 이용한 순수한 이용자 수를 나타내는 지표
- conversion(전환) : 고객이 디지털 마케팅의 영향을 받아 구매나 그에 가까운 행동을 하는 것
- 전환율 : 특정 인터넷 서비스에 방문한 사람 중, 해당 서비스에서 유도된 행위를 한 방문자의 비율
- 정규화 : 데이터의 평균을 0으로 한 후 평균에서 어느 정도 떨어졌는지를 분포화해서 나타내는 방법
- 인포그래픽 : 정보가 빠르고 분명하게 표현하기 위해 정보, 자료, 지식을 그래픽 시각적으로 표현한 것을 의미한다. (예: 차트, 사실박스, 지도, 다이어그램, 흐름도 등)