20230412

1. 이진 데이터란?

- 브리스톨 병원의 예 (PPDAC) ⇒ Problem : - 브리스톨 병원의 어린이 심장 수술 사망률은 다른 병원과 비교해서 현저히 높음?
  - 어린이란?
  - 심장 수술이란?
  - 사망의 기준은?

plan : - 심장 수술 database → 비교

Data : HES, CSR 이라는 자료 사용. but, 개복수술 사건과 사망건수가 다름

Analysis : HES, CSR 2의 혼합자료 사용 → 사망률 예측

Conclusion : 예측 기준 3?명이지만, 실제 사망자는 62명

- binary data(이진 데이터) : 두 가지 값으로 이루어진 데이터. (ex. 1=사망, 0=생존)
  - 평균은 비율로 표시된다. (ex. 사망률)

- '결론' 보다 중요한 것은 효과적으로 전달 (ex. 사망률 ⇒ 부정 메시지 프레이밍
  생존률 ⇒ 긍정 메시지 프레이밍)
  ∴ 긍/부정 프레이밍, 절대-상대적인 요약 모두 활용하여 올바르게 정보 전달 해야함!
  - 막대 그래프의 기준점을 합리적으로 세워야함.

2. 범주형 자료의 소개

- 범주형 자료란? : - Variable : 주어진 상황에 따라 다른 값을 가지는 측정치
  └> 2개 이상의 범주를 값으로 가지는 변수
  - 순서 X 범주 : 국적, 성별
  - 순서 o 범주 : 연령대
  - 일련 2분 숫자들 : BMI 기준 비만 측도

  ※ 파이차트는 절대 사용하지 마라!!
  (∵ 숫자를 봐야만 알 수 있는 시각화 자료임. ⇒ 시각화 능력이 떨어짐)
  → 대신 막대 그래프를 사용하라.

- 상대위험도 ⟷ 절대 위험도
  6%, 7%

위험요인이 있는 질병위험도

Control group 질병위험도

∴ 상대위험도가 높다고 절대위험도 자체가 높은 경향에는 큰 문제가 아닐 수 있다.

- 기대도수 : 주어진 집합에서 특정 사건이 일어날 개수의 예측값
- 오즈 : 도박에서 많이 사용됨 (ex. 한국 월드컵 우승 확률이 $\frac{1}{30}$인 경우 $\frac{1/30}{1-1/30} = \frac{1}{29}$ 우승의 경우의 수 / 우승×경우의 수)

- 오즈비 = 위험요인이 있는 집단의 오즈 / Control group의 오즈 # 절대위험도가 작을 경우 오즈비 ≒ 상대 위험도

3. 연속형 자료와 요약

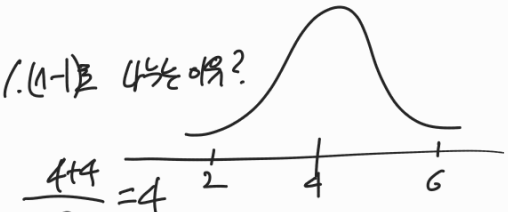
- 범주형
  - 순서 O
  - 순서 X
- 수치형
  - 이산형 : 값이 정수
  - 연속형 : 값이 실수

- 연속형 자료의 시각화
  - step-chart
  - box plot
  - 히스토그램

- 데이터 처리법과 표준화

- 자료의 위치 대표값
  - 중앙
    - 평균 $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
    - 중앙값
    - 최빈수
  - 퍼짐
    - 범위
    - IQR
    - 표준편차
    - 분산 $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

Q) 1. (n-1)로 나누는 이유?



$\frac{4+4}{2} = 4$

- 자유도

2. 제곱대신 절대값을 사용하면?
- 미분관점으로 전개하면 어려움이 있었음

- 산포도 통계량: 한 개의 데이터 집의 변화 크게 차가 적지 않는 통계량
  ex. IQR, 중앙값

- Strip chart (dot plot): 자료들이 겹쳐있는 점의 색이 진해짐

- 히스토그램: 구간으로 나누어 그림 ∴ 구간의 개수, 크기, 시작점이 중요함.

- Modality ⟷ Skewness
  ⇒ 봉우리가 몇개냐
  ex. 히스토그램
  ⇒ 대칭적이냐