[2주차]패턴인식과 패턴탐색

패턴탐색

1. **확증적 자료 분석(Confirmatory Data Analysis)**

   추론통계 : 데이터를 이용하여 모집단의 파라미터와 분포를 추정하여 모델기반에서 예측하는 기법

2. **탐색적 자료 분석(Exploratory Data Analysis)**

   기술통계 : 데이터를 요약 설명하는 통계기법으로 데이터의 중심, 산포, 상대적 위치, 그리고 분포 등을 활용

   → 탐색적 자료 분석은 원데이터에 대한 탐색과 이해의 첫 단추

   1. **모집단의 대푯값**

      자료의 중심 : 평균, 중앙값, 최빈수

      자료의 상대적 위치 : 사분위수, 백분위수

      자료의 산포 : 평균 절대 편차(MAD), 사분위수 범위(IQR), 분산, 표준편차

      탐색적 자료 분석에서 데이터의 패턴을 본다 → 산포를 볼 수 있는 측정량(자료의 산포)이 중요

      상관계수 : 두 특성의 변화 관계

   2. **모집단과 표본의 대푯값** : **자료의 산포**

      a. 모집단의 산포 측정 모수 : 모분산

      b. 표본의 산포 측정 통계량 : 표본분산

      표분분산이 모집단의 산포정도를 추정하는 통계량으로 사용되는 이유 : 표분분산이 가지는 불편성(unbiased) 때문

      모집단의 산포도를 알 수 있는 모분산을 표본을 통해 추정해야하는데 표본분산이 얼마나 모분산을 오차없이 추정할 수 있는가 → 불편성을 만족하기 위해 n-1로 Scaling 조정 → 모집단의 모분산을 표본분산으로 설명을 잘한다.

자료분석의 통계적 접근

1. 문제제기(What to measures?)에 해당하는 목표모집단 설정
2. 모집단을 잘 닮도록 확률표본 설계 및 데이터 수집

3. 데이터시각화, 변환, 빅데이터기초분석 및 모델링을 통한 문제제기에 솔루션 제공
4. 재생산적(Reproductional) 문제제기 창출

→ 이러한 과정을 패턴 탐색이라는 분류에 속함

자료분석의 통계적 접근의 활용 사례

제조업

공정관리에서 발생하는 데이터를 분석하여 불량품의 원인을 규명, 예방하는 품질관리(Quality Control)에 활용

금융분야

고객의 신용 등급에 따라 대출 규모와 이자 등을 결정하는 신용점수(Credit Score) 산정에 활용

부정행위적발

특이한 거래 행위에서 부정행위를 적발(Fraud detection)하는 분야에 활용
- 잃어버린 신용카드의 부정이용
- 보험회사의 허위과다 청구를 예방하기 위해 사용
- 국민연금이나 의료보험의 부당 창구와 같은 영역에서도 활용

데이터 측정

측정의 타당성

문제적 정의와 조사한 단어의 기준을 파악하여야 함

⇒ 통계적 목적에 부합하도록 측정하고자 하는 특성을 정의하는 최선의 방법
은 그것을 측정하는 규정을 타당성 있게 정의하는 것

특성을 측정한다는 것은 그 특성을 나타내는 방법으로 각 단위에게 수치를 부여함을 뜻함

편의(bias)와 정도의 결여(lack of precision)의 판단 필요

데이터의 이해

1. 데이터의 출처는 무엇인가?

   자료의 신뢰성 → 올바른 이용여부 조사 → 자료의 질 평가
2. 이치에 맞는 데이터 인가? → 내적 일치성을 살핌 → 숫자들 사이에 일관성이 있는가?
3. 데이터는 완벽한가?

4. 잘못된 계산은 없는가?

데이터의 내적, 외적 타당성

숫자를 통하여 통찰력을 제공

문제에 대해 유효하고 관련이 있는 타당한 데이터 수집

⇒ 내적타당성, 외적타당성

내적타당성

조사나 실험 대상과 관련

타당한 측정여부

외생 변수와의 교락, 계산 실수나 불완전한 정보

외적타당성

결론이 더 큰 모집단으로 일반화될 수 있는가?(외적으로 확장하는게 가능한가)

실험에서 비현실적 처리나 모집단을 대표하지 못하는 조사대상으로 부터도 정확한 데이터는 얻을수 있지만 결론은 얻을 수 없다.

John W. Tukey ← Data Analysis 의 기틀

Approximate solution to a right problem is better than an exact solution to a wrong problem

→ 문제제기(문제파악)의 중요성을 지적

→ 단, 관찰에 의한 조사연구에서는 인과관계를 유도할 수 없으며, 그 조사시점과 상황에서의 실태파악만 가능함

Ex) 코로나 백신으로 사망한 사람들이 백신이라고 하는 인과성을 밝힐수 없다 라고 하는것은 관찰에 의해서만 얻어질 수 밖에 없기 때문

동일한 조건내 control할 수 없기때문에 어떤 원인인지 밝힐수 없는 것

백신이 원인중에 하나일수는 있지만 백신만이 원인은 아님

수집된 데이터가 측정과 관찰에 의해서 얻어진 데이터라면 분석과 관계에서 절대로 인과성을 이야기 할 수는 없음