금융인을 위한

통계와 데이터 분석 입문

통계적 오류 피하기

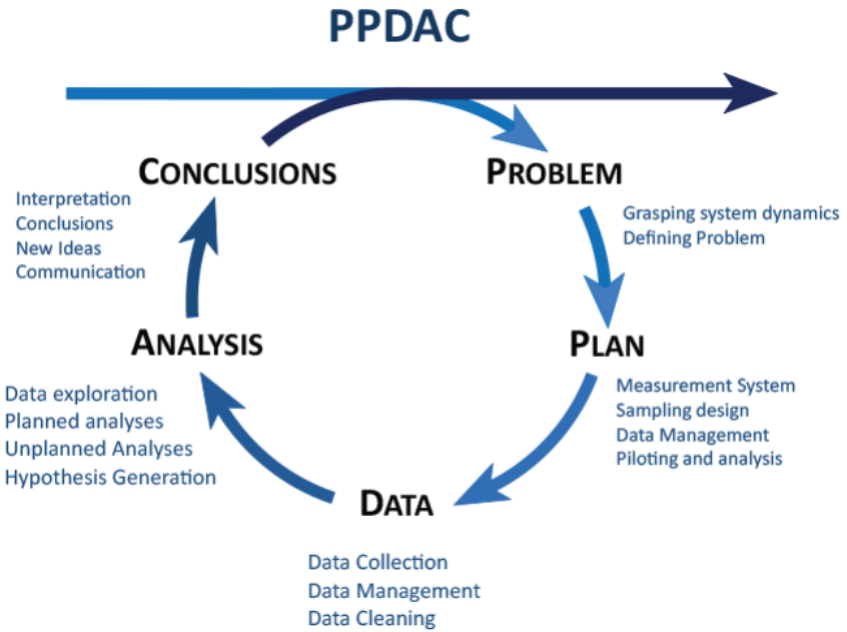
한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 데이터분석 단계별 통계오용 및 왜곡
2. 통계적 주장에 대한 검토

데이터분석 단계

- 데이터분석 PPDAC의 각 단계에서 통계적 오류가 발생할 가능성이 존재함



출처	Problem, Plan, Data, Analysis, and Conclusion (PPDAC) Model (Wild and Pfannkuch, 1999)

통계의 오용 · 왜곡 예시

예

월급쟁이 평균소득 297만원…남성소득은 여성의 1.5배

출처 통계청

- 전세계적으로 성별에 따른 임금격차는 성차별 주요 요인으로 인식되고 있으며 사회적 이슈 중 하나임

통계의 오용 · 왜곡 예시

예

월급쟁이 평균소득 297만원…남성소득은 여성의 1.5배

- 문제 1. 월평균 소득 같이 이상치가 존재하는 데이터의 경우 평균은 대표성을 갖지 못할 수 있음
  - 이상치를 제거하고 계산된 절사평균이나 중위수 활용 고려
- 문제 2. 임금 차이에 영향을 줄 수 있는 다른 변인들을 고려하지 않았음
  - 교란변수(confounding variable)들을 고려할 필요 : 예를 들어, 노동시간, 근속연수, 연령에 따른 경제활동 인구 등

통계의 오용 · 왜곡 예시

예

교육수준 낮은 여성 '대사증후군' 많아…남성은 차이 없어

출처 국민건강영양조사 (2015), YTN

- 대사증후군의 원인으로 비만과 나쁜 식습관 등이 있는데 여성의 경우 학력이 높을수록 체형 관리에 신경을 쓰기 때문에 대사증후군의 발병이 낮아진다는 해석

통계의 오용 · 왜곡 예시

예

교육수준 낮은 여성 '대사증후군' 많아…남성은 차이 없어

- 문제. 여성의 경우 학력 등 사회경제적 지위와 발병률이 강한 상관관계를 보였다고 하였는데 연령대와 학력수준과의 관계를 제대로 고려하지 않음
  - 발병률을 연령대별로 나눠서 남녀를 비교해본 결과 큰 차이가 나오지 않음
  - 위와 같은 결과는 20-40대 여성의 학력수준 증가가 원인일 수 있음
  - 학력 뿐만 아니라 연령대에 따른 적절한 표본추출과 학력과 연령대 간의 관계 분석 필요
- 문제. 강한 상관관계를 인과관계로 확대 해석하는 것을 피해야 함

통계의 오용 · 왜곡 예시

예

층간 흡연에 대한 불편함을 느끼지 않다는 사람의 비중이 높았던 조사 결과
조사대상 : 전국 만 19세 이상 남녀 1,000명 (유선 조사)
응답률 : 22%

- 문제 1. 조사대상 연령에 대한 제한
  - 층간 흡연에 대한 의견은 그 이하 연령층에서도 낼 수 있으므로 조사대상을 확대
- 문제 2. 유선전화 조사의 문제
  - 유선전화 조사의 경우는 낮은 응답률을 가지고 있을 뿐만 아니라 휴대전화만 보유하고 있는 사람은 조사에서 배제되므로 조사방법 수정

통계의 오용 · 왜곡 예시

예

층간 흡연에 대한 불편함을 느끼지 않다는 사람의 비중이 높았던 조사 결과
조사대상 : 전국 만 19세 이상 남녀 1,000명 (유선 조사)
응답률 : 22%

- 문제 3. 조사대상의 흡연 여부 고려
  - 본인의 흡연 여부에 따라 불편함을 느끼는 여부가 달라질 수 있으므로 흡연 여부를 조사하고 이에 따라 결과를 해석
  - 흡연여부, 연령대 등을 조사에 포함하여 결과에 영향을 줄 수 있는 요인 통제

계획 · 데이터 수집 단계에서의 오류

- 편의와 비용을 위해 **대표성**이 부족한 표본을 선택한 경우
- 설문조사에서 특정 선택을 유도하거나 오해를 불러 일으키는 문구를 사용한 경우
- 자원한 사람들만 가지고 연구하는 등 공정하지 못한 표본으로 비교한 경우
- 표본의 크기가 너무 작거나 응답누락, 중도탈락이 많은 경우
- 잠재적인 교란변수를 잡아낼 수 있는 데이터 계획에 실패한 경우

예

운동하는 양 - 물 섭취량 - 비만율

→ 계획 단계에서부터 통계전문가가 투입되어야 함
: '사후부검'에 비유함

데이터 분석 단계에서의 오류

- 계산, 코드의 실수
- 결론을 맞추기 위한 데이터 정리

예

‘A 후보와 B 후보 중 A 후보가 더 좋다고 생각하십니까?’ 의 질문에 대한 대답을 ‘그렇다’, ‘보통이다’, ‘아니다’로 하게 한 경우

데이터 분석 단계에서의 오류

- 통계적 방법의 오류 및 잘못된 추론
  - 여러 복잡한 요인을 간과한 단순한 모형 적용 및 해석
  - 모형에 대한 가정 위반
  - 동등하지 않은 그룹 비교, 잘못된 기준 사용 등

예

고용형태(정규직, 비정규직)에 따른 근로조건 차이를 보는데 있어서 비정규직에 근로시간이 짧은 시간제근로자를 포함시켜 단순 비교

결론 보고 단계에서의 오류

- 숫자와 비율 등을 과장해서 눈속임하는 경우

예

1시간 → 2시간 : 200% 증가 vs. 50% 감소

- 통계적 검정을 여러 개 수행한 후 가장 유의미한 결과들만을 발표하는 것

예

미국의 한 제약회사에서 부분집합 분석의 유의미한 결과들만을 선별적으로 발표

→ 유죄판결

통계적 주장에 대한 검토

- 데이터에 기반한 통계적 주장은 다음 조건을 충족해야 함
  - 독자는 정보에 접근 가능해야 함
  - 독자는 정보를 이해할 수 있어야 함
  - 독자가 원한다면 주장의 신빙성을 확인, 평가할 수 있어야 함
  - 독자가 정보를 활용할 수 있어야 함

통계적 주장에 대한 검토

- 제시된 숫자들에 대한 검토
  - 적절한 설계, 질문의 단어선택, 실험 지침에 대한 사전 등록, 표본의 대표성, 무작위 배정, 공정한 비교 등을 점검함
  - 자료의 요약에 평균, 변동성 등 기초통계량이 적절하게 사용되었는지를 점검함
  - 결과에 대해 신뢰구간, 통계적 유의성, 표본 크기, 다중 비교 등을 점검함

통계적 주장에 대한 검토

- 출처에 대한 검토
  - 통계가 참고한 출처에 대해 점검함
  - 극단적인 사례들에 대한 인용이 있는지,
    오해를 불러 일으키는 그래프·과장된 헤드라인 등이
    사용되었는지를 점검함
  - 통계에서 보여주지 않은 것은 어떤 것들인지에 대해 점검함
    : 선별적 발표에 대한 경계

통계적 주장에 대한 검토

- 해석에 대한 검토
  - 적절한 비교가 이루어졌는지, 과거 다른 연구 결과들과 상충되는지에 대해 점검함
  - 통계 분석의 결과를 바탕으로 내릴 수 있는 결론인지, 확대 해석된 것은 아닌지를 점검함

예

상관관계 vs. 인과관계, 유의하지 않음 vs. 효과 없음

  - 일반화의 오류가 있는지를 점검함

→ 통계적 주장에 대한 검토를 통해 질 낮은 통계분석을 언제나 경계해야 함

학습정리

- **통계적 분석 단계**는 PPDAC(problem-plan-data-analysis-conclusion)으로 정리해볼 수 있음
- 통계적 오류는 데이터 분석의 **모든단계**에서 발생할 수 있음
- 통계적 주장에 대한 **검토**를 **숫자, 출처, 해석 등 다양한 각도**에서 함으로써 질 낮은 통계분석을 언제나 경계해야 함