

# 제 1 장 서 론

## 1. 탐색적 자료분석 - EDA-

◆ EDA(Exploratory Data Analysis) 란?

데이터의 특징과 내재하는 구조적관계를 알아내기 위한 기법들을 총칭

◆ 자료분석의 제단계

자료분석 ➡ 탐색단계와 확증단계로 구분

탐색적 자료분석(EDA)

데이터의 구조와 특징을 파악하며 여기서 얻은 정보를 바탕으로 통계모형을 만드는 단계

확증적 자료분석(CDA : Confirmatory Data Analysis)

관측된 형태나 효과의 재현성 평가, 유의성 검정, 신뢰구간 추정 등 통계적 추론을 하는 단계

(1) 관련된 다른 자료분석에서 얻어진 정보를 적절히 배려하는 일

(2) 새로 수집된 자료가 앞서의 분석결과에 의한 예측과 얼마나 일치하는가를 평가하는 일

## EDA의 역사

존 튜키(John W. Tukey : 1977년)의 첫 저서 “탐색적 자료분석(EDA)”  
자료가 무엇을 말하려 하는가를 보기 위한 것

단순한 계산과 그리기 쉬운 그림에 집중

추론통계학이 아닌 **기술통계학**(descriptive statistics)의 중요성을 강조

## 2. 탐색적 자료분석의 네 가지 주제

- EDA의 네 가지 주제

1. 저항성(resistance)의 강조
2. 잔차(residual) 계산
3. 자료변수의 재표현(변수변환 : re-expression)을 통한 다각적 시도
4. 그래프를 통한 현시성(revelation)

## 1) 저항성의 강조

탐색적 자료분석은 일부자료의 파손에 관한 저항성을 가져야 한다.

05, 08, 09, 10, 11, 15, 16, 19, 21, 29, 31
--

산술평균 (Mean) = 15.8

중위수 (median) = 15

# 자료의 마지막 값 31이 310으로 변경된 경우

산술평균 (Mean) = 41.2

중위수 (median) = 15

# 문제의 관측값 31을 x로 대체되는 경우

산술평균(x) (Mean(x)) =  $(143+x)/11 = 13+(1/11)x$


중위수(x) (Median(x)) =  $\begin{cases} 11, & x \leq 11 \text{인 경우} \\ x, & 11 < x \leq 15 \text{인 경우} \\ 15, & x > 15 \text{인 경우} \end{cases}$

\*\*\* EDA의 관점에서는 평균보다는 일부자료의 파손(변형)에

저항적인 **중위수**가 바람직한 **대표값의 척도로 선호된다.**

## 2) 잔차 계산

잔차는 각 개별 관측값이 자료의 주경향(main trend)으로 부터 얼마나 벗어났는지를 나타냄  
중위수 15로부터의 잔차



05, 08, 09, 10, 11, 15, 16, 19, 21, 29, 310
-10, -7, -6, -5, -4, 0, 1, 4, 6, 14, 295

왜 이렇게 큰 잔차 295가 주어졌는가 의문을 풀어보자 (EDA의 업장)

## 3) 자료변수의 재표현(변수변환) →

자료의 변환으로(측정척도를 적당히 다른 척도로 재표현) 분포의 대칭성,  
관계의 선형성(직선화), 분산의 균일성, 관련변수의 가법성등에 도움이 된  
다.

## 4) 그래프를 이용한 현시성(revelation)

그래프 표현이 자료안에 숨겨진 정보를 보여주는 효율적인 수단이 된다.  
탐색적 자료분석에서는 다양한 그래프 작성 기법들이 사용된다.

수치 데이터 ← 추상적인 개념

구상화

각종 그래프

도표1 수리통계학 점수 자료

A	XX	A	XX	A	XX	A	XX
0	54	1	67	1	55	0	23
0	51	0	64	1	90	1	51
1	52	0	43	0	15	0	10
0	82	1	74	1	54	1	78
1	37	0	73	0	52	0	48
1	41	0	33	0	52	0	30
1	41	1	51	1	18	1	39
1	46	1	28	0	53	1	44
1	46	1	56	1	28	1	58
0	29	1	58	1	67	0	35
0	25	1	38	1	61	0	53
0	23	1	73	1	69	0	47
0	41	1	45	1	77	0	56
0	89	0	28	0	54	1	99
0	0	0	43	0	35	0	24
0	21	0	23	1	67	0	0
0	53						

과제 미제출그룹 (A=0)

$\frac{1}{\sqrt{x}}$

중위수=42  
50점대와  
20점대의  
혼합모형

과제<sub>1</sub> 제출그룹 (A=1)

1771

중위수  
= 54

54

#### 4. 데이터 분석의 기본철학

과학발전에 있어서 통계학(또는 통계 전문가)의 역할은 무엇인가?

##### 1) 모형-데이터의 사이클

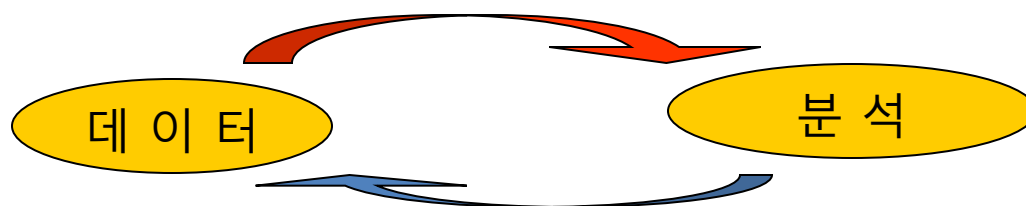
전통적인 통계학  
가설검정) ← 모형이 주어진 데이터의 분석 (추정 또는  
실제적인 연구시



##### 2) 진실로서의 모형, 대표적 경우로서의 모형

통계적 모형은 진실로서의 모형이 아니라 대표적인 경우로서의 의미가 있을 뿐이다  
(예 : 회귀분석에서 오차항의 정규성 가정)

##### 3) 데이터에 대한 올바른 이해



## 5. 탐색적 데이터 분석의 성공사례

사례)

1973년 미국 뉴저지 주의 주지사는 주정부 환경보호과로부터 대기 중의 오존 수준을 안전 수준으로 낮추어야 하며 그러기 위해서는 자동차 배출가스의 총량을 3분의 2로 줄여야 한다는 내용의 건의

문제점)

- 1) 일요일에 교통량이 적음에도 불구하고 오존수준은 요일별로 차이가 없다
- 2) 가장 높은 오존수준을 나타내는 곳은 농촌지역인 양코라였다  
측정장치의 소홀한 관리와 영성한 보정이 원인

연구에 중요한 요소

- 1) 자료를 수집하고 분석에 임하는 열의
- 2) 기대하지 않았던 사실을 발견하기 위한 진단기법의 활용
- 3) 놀라운 사실을 인지할 수 있는 능력과 노력
- 4) 해당분야에 대한 충분한 이해와 해당분야 전문가와의 커뮤니케이션
- 5) 자료의 효과적인 요약과 그래프의 작성
- 6) 연구의 실마리에 대한 끊임없는 추적