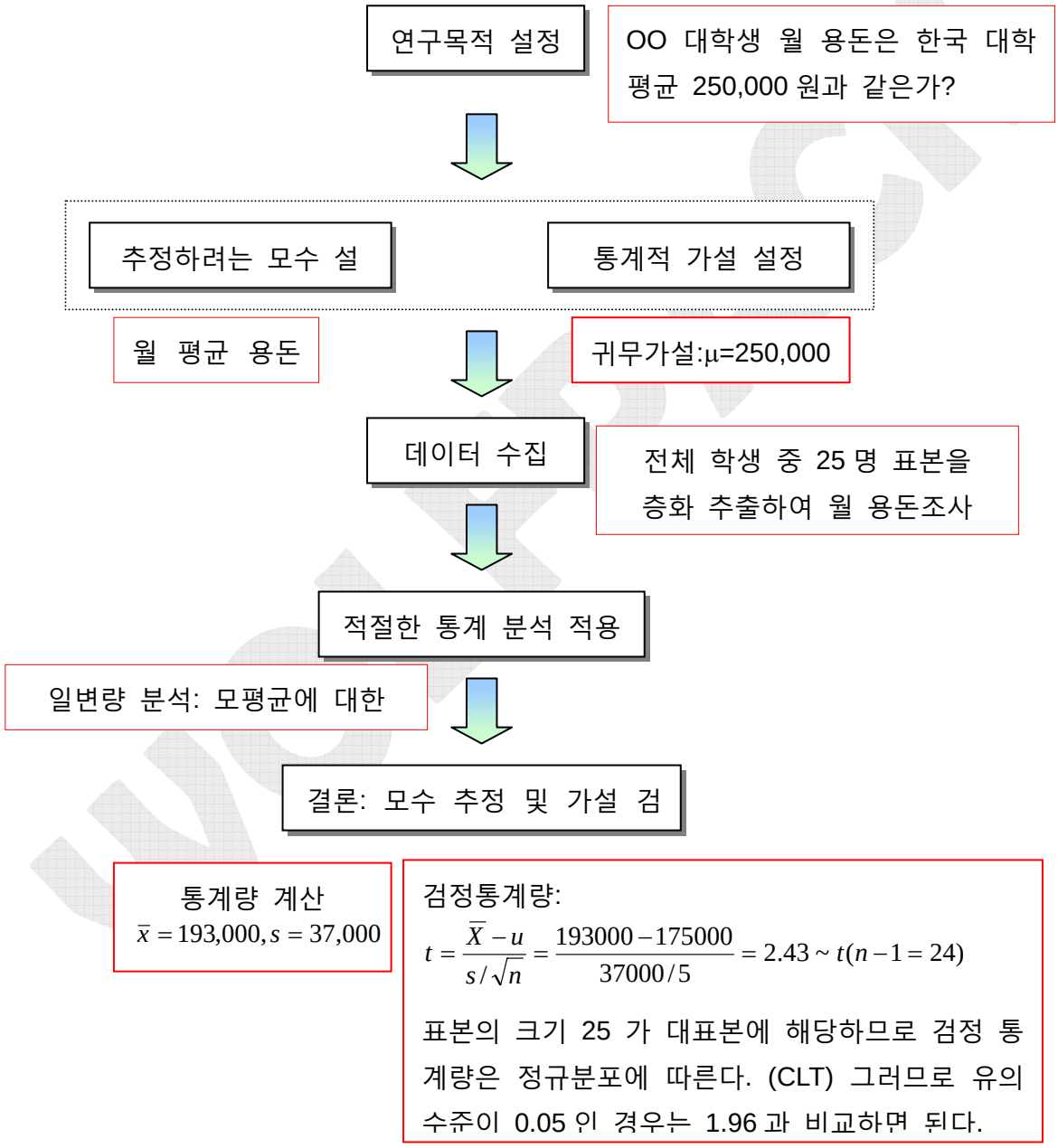
고전적 데이터 분석은 연구 목적이 설정되면 그에 맞는 1)통계적 가설(statistical hypothesis), 모형(model)을 설정하고 2)데이터 수집하여 3)가설 혹은 모형의 유의성(significance)을 검정하였다. 이를 Confirmatory (확증적) Data Analysis 라 한다.
다음은 (confirmatory) 데이터 분석의 예로 한남대 학생들의 용돈이 대학 평균과 같은가를 알아보는 연구 과정을 요약한 것이다.



1.1 EDA 정의

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights. Its concern is with appearance, nor with confirmation. - Exploratory Data Analysis, John W. Tukey, 1977 -

CDA (확중적 데이터분석)와는 달리 1977 년 John W. Tukey 가 (Princeton University Bell Lab) 제안한 탐색적 데이터분석 (EDA: Exploratory Data Analysis) 방법은 이미 수집된 데이터로부터 정보를 얻어내는 일련의 방법이다.

1) 이미 수집된 데이터가 가진 정보를 간편한 계산식에 의해 구해진 숫자 요약(중앙값, 사분위)과 그래프(예: stem-leaf plot, box plot, scatter plot)를 이용하여 찾아내거나
2) 데이터를 보다 유용하게(정규분포 혹은 대칭인 분포) 만들기 위하여 데이터를 재표현(re-expression = data transformation 예: log 변환) 하거나
3) 데이터가 어떤 분포에 적합한지 알아보는 방법에 관련된 데이터 분석 방법이다. (적합성 검정)

1) EDA is about looking at data to see what it seems to say
데이터가 가진 정보를 데이터의 탐색만으로 얻는 방법이다. 이전 통계학이 추론 통계에 의존했다면 EDA 는 통계학이 기술 통계학 (descriptive statistics)임을 강조하고 있고 통계적 가설 설정 과정이 없다.

2) EDA is a detective work.
여러 도구 (tools: 기술통계량, 관련 그림)와 직감(intuition: 데이터 분석 경험에서 얻는 분석 know-how) 이용하여 정보(결론)를 유추하는 분석 방법이다. CDA 는 판사의 (judge) 작업이라면 EDA 는 여러 정황을 고려하여 사건을 분석하는 탐정과 같은 역할이다.

3) To learn about data analysis, it is right that each of us try many things that do not work.
데이터로부터 정보를 얻기 위한 다양한 시도를 해야 한다. 데이터를 다루는 풍부한 경험(비록 성공하지 못하더라도)으로부터 올바른 데이터분석이 가능하기 때문이다.

4) EDA can never be the whole story, but nothing else can be served as the first step.
탐색적 데이터 분석은 분석의 모든 것은 아니지만 첫 단계가 된다. 탐색적 데이터분석을 통해 얻은 정보를 이용하여 통계적 가설이나 모형을 설정하여 연구하기도 하고 의사 결정에 이용하여 정보의 정확 정도를 측정하기도 한다.

5) EDA is a paper-pencil method.
컴퓨터(소프트웨어)가 보편화 되지 못하고 데이터의 수가 적을 때 그래프나 통계량들을 직접 그리거나 계산하기에 편리하게 제안된 방법이기 때문에 이런 별명을 가지고 있다. 요즈음은 통계 소프트웨어의 발달로 쉽게 그리거나 구할 수 있으므로 정보 얻는 방법, 해석 방법을 이해하는 것이 중요하다.

6) Data Mining is a modern EDA.
신용 카드, 멤버쉽 카드, 교통 카드 등 카드 사용에 의해 데이터가 자동 수집되고 OLTP(On-Line Transaction Process: 데이터 자동 수집) 수집된 데이터를 잘 저장하는 방법 Data Warehousing 기술의 발달로 Data Mining (대용량의 데이터에 내재되어 있는 patterns 이나 rules 을 발견하는 방법)이 가능해졌다. Data Mining 도 일종의 EDA 이다. Data Mining 으로부터 얻은 정보를 이용하여 고객관리 하는 방법을 CRM 이라 (customer relationship management) 한다.

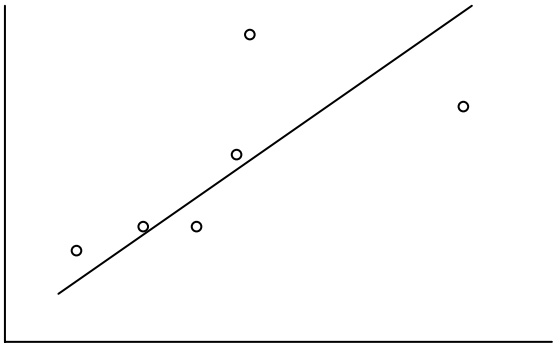
1.2 EDA 4 가지 주제

1) Resistance to outliers, missing data, or miscoded data

이상치, 결측치, 입력 오류에 영향을 받지 않는(resistant) 도구를 사용한다. EDA 에서 수집 데이터의 숫자 요약 통계량으로 중앙값, 사분위수 등을 이용하는 이유이다.
(예) 1 2 3 4 10 ➔ 평균=5 그러나 중앙값은 3 이 된다.

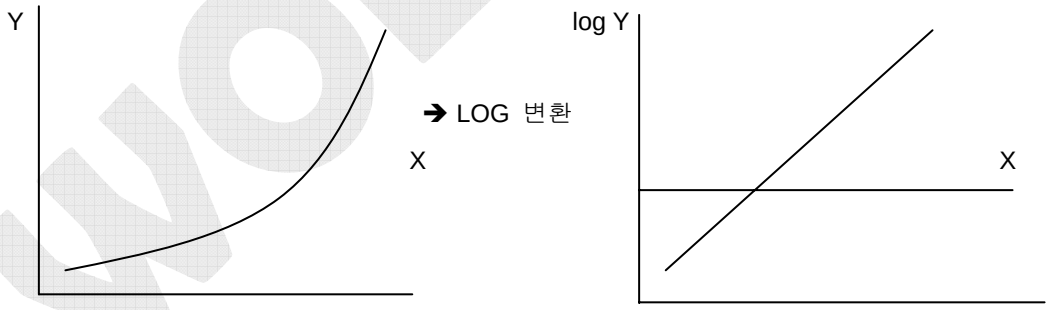
2) Residual is a off-value from the main stream

잔차는 각 값들이 주 경향으로부터 얼마나 벗어나 있는지 나타내는 값이다. 앞의 예에서 중앙값을 중앙(main stream)으로 사용하는 경우 잔차는 −2 −1 0 1 7 이다. 그러므로 마지막 값에 대해 왜 이런 일이 발생했는지 탐색 작업이 필요하다. 다른 예를 살펴 보면 회귀 분석에서 직선의 경향이 벗어난 관측치가 이상치(outlier)인지 영향치(influential)인지를 산점도를 이용해 판단할 수 있다. 이상치나 영향치나 모두 잔차(추정 회귀 직선에서 벗어난 정도)가 크다는 공통점은 있으나 영향치는 이상치와는 달리 다른 관측치에 비해 이상할 정도로 벗어나 있다는 근거를 제시할 수 없는 관측치를 일컫는다.



3) Data Re-expression

원래 데이터를 Log(로그), Square root(제곱근), Inverse(역) 변환 등으로 데이터 값을 변화시키는 것을 데이터 재표현이라 한다. 이는 데이터의 분포의 정규성(아니 엄밀히 말하면 대칭성), 균일성(uniformity), 가법성(additivity)을 얻기 위하여 시행한다. 통계 데이터 분석 기법의 대부분은 변수의 정규성(적어도 대칭성)을 가정하고 있다. 예를 들어 페이지 1 에서 표본을 25 명이 아니라 15 명만 뽑았다면 검정통계량은 더 이상 정규분포를 따르지 않는다. (즉 CLT: 중심 극한 정리) 이런 경우 모집단은 정규 분포를 따른다는 가정이 있어야 t-분포를 이용할 수 있다. 만약 모집단이 정규분포를 따르지 않는다면 데이터 재표현(변수 변환)을 통해 데이터가 정규성을 만족하게 하여야 한다.
다음은 두 변수간의 관계를 나타낸 그래프, 즉 산점도(scatter plot)이다. 왼쪽 산점도에 의하면 Y 와 X 의 관계는 직선관계가 아니다. 대신 Y 를 재표현(변수 변환) 하여 LogY 와 X 에 대한 산점도(오른쪽)를 그리면 직선 관계가 존재한다. (직선 관계를 분석하는 것이 결과 해석이 편리)



4) Graphic presentation

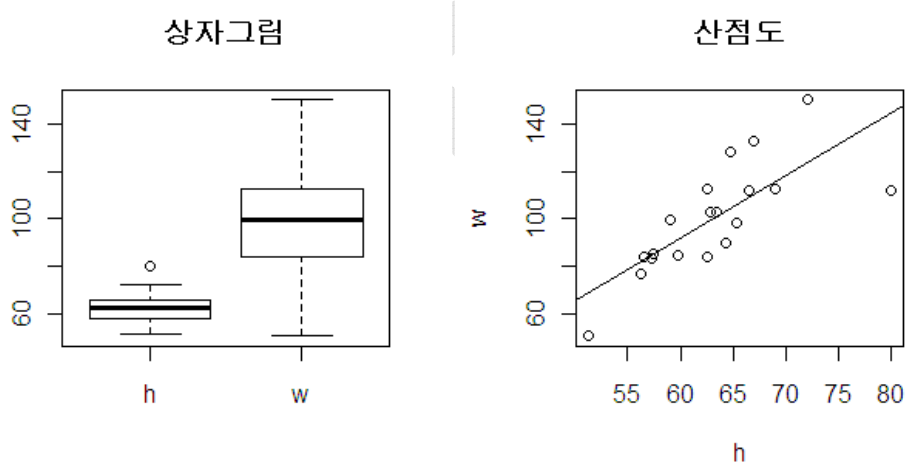
EDA 에서는 데이터에 숨겨진 정보를 알아보기 위하여 다양한 그래프가 이용된다. 다음은 키 데이터에 (변수) 대한 줄기-잎 그림, 상자 그림과 키와 몸무게의 관계를 나타내는 산점도를 그린 예이다. (SAS Example Data)

그래픽 표현

SAS CLASS 데이터 중 키의 (단위: inch) 마지막 데이터를 '80'으로 수정하였음. 몸무게 (단위: pound) 데이터는 동일.

```
> h=c(69,56.5,65.3,62.8,63.5,57.3,59.8,62.5,62.5,
+ +59,51.3,64.3,56.3,66.5,72,64.8,67,57.5,80)
> w=c(112.5,84,98,102.5,102.5,83,84.5,112.5,84,
+ +99.5,50.5,90,77,112,150,128,133,85,112)
> ds=data.frame(cbind(h,w))
> boxplot(ds,main='상자그림')
> plot(h,w,main='산점도')
> abline(lm(w~h))
```

- (상자수염 그림) 키의 경우 이상치 하나 존재, 몸무게의 흩어짐 정도가 큼, 좌우 대칭 분포 형태를 갖는다. 이상치 제외하면 모수적 데이터 분석에 문제 없음
- (산점도) 키와 몸무게 간에는 직선적 관계가 존재, 이상치 하나 있음



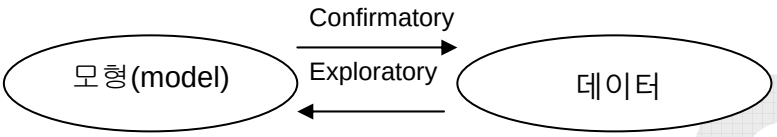
1.3 데이터 분석의 기본 철학

"탐색적 데이터 분석" 허명회 & 문숭호, 자유아카데미, 2000

과학은 이론적 통찰 (예: 상대성 이론), 새로운 현상의 관찰 (Kepler 행성 궤도 관련 법칙)이나 경험을 (Student T-분포) 통한 새롭고 혁신적인 이론이 만들어지는 경우는 극히 드물고 대부분 관찰, 실험, 분석 등의 반복을 통해 이론이 정립된다. 벼 품종 개량, 새 의약품 개발, 화학 공정 개선 등이 실험 계획에 의한 연구 결과가 이에 해당된다.

통계 전문가는 제시된 이론을 통계적 가설이나 통계 모형으로 설정하고 관련 데이터를 수집하여 가설(모형)의 유의성을 검정하거나(confirmatory data analysis) 수집된 데이터를 탐색하여 가능한 모형이나 이론을 제시하는 역할을(exploratory data analysis) 담당하고 있다. 이처럼 (탐색적) 데이터 분석이 타 분야의 새로운 이론 발견에 기여할 수 있으려면 1)그 분야에 대한 지식 2)모형과 데이터 3)그리고 모형과 데이터의 사이클 개념을 올바로 이해해야 한다.

1.3.1 모형과 데이터 사이클



$f(x:\theta)$ 모수 $\qquad (x_1, x_2, ..., x_n)$

$y = f(x_1, x_2, ..., x_p)$ 함수 $\qquad (y_i, x_{1i}, x_{2i}, ..., x_{ni})$, $i = 1,2,..., p$

과학에서 이론이 제안되고 데이터 분석이 이루어지는 경우보다는 데이터로부터 새로운 이론이나 모형을 도출하는 경우가 많고 탐색적 자료 분석에 의해 제안된 이론이나 모형은 다시 confirmatory 방법에 의해 유의성이 (significance) 검증되므로 모형과 데이터는 순환 사이클을 갖는다.

통계적 모형은 과학적 진실이기 보다는 사실의 대표적 모형이다. 예를 들어, 회귀모형에서는 ($y = a + bx + e$) 설명되어지지 않는 오차항이 존재하고 이 오차항은 $iid \sim N(0, \sigma^2)$ 을 가정한다.

1.3.2 탐색적 데이터 분석의 성공 사례

1973 년 미국 뉴저지 주지사는 오존 수준을 안전 수준으로 낮추기 위하여 자동차 배기 가스를 현재 수준의 2/3 으로 줄이는 법안 입안을 요청 받았다. 이 법안의 타당성 조사를 벨 연구소(Bell Lab)에 의뢰하였다. 7 년간 60 개 측정소에서 300 만개 측정 자료를 수집하여 plot 한 결과 1)최고 오존 수준은 요일별 차이가 없고 => 원인 규명이 어려움 2) 농촌 지역인 Ancora 지역에서 높은 오존 수준 보인다는 특이한 사실을 발견하였다. 2)의 원인으로 이 지역에서 37km 떨어진 Philadelphia 지역의 공해 물질이 바람에 날려와서 오존 수준을 높였을 가능성이 주장되었다. 이 주장은 오존 수준과 풍향과의 plot 을 통해 사실임이 밝혀졌다.