효과적인 통계분석을 위한 기본 과정

“세상의 이치, 숫자 속에 그 답이 있다.”

정수진

목차

1. 통계학개론

1) 통계란?

한스 로슬링(Hans Rosling)

- 스웨덴 카롤린스카 연구소 교수
- 갭마인더 재단 대표
- 의학통계, 공중위생학 전공

*** 통계 없이 세상을 이해할 수 없다 ***

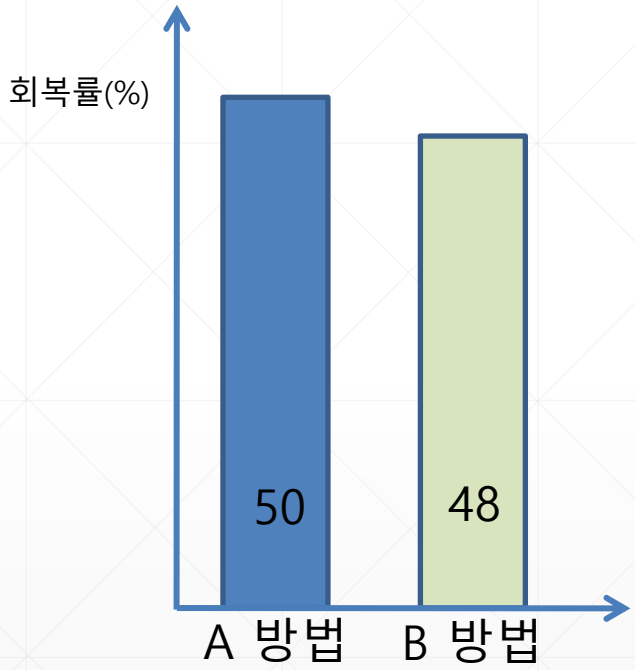
✓ 우리 주위의 현상에서 데이터를 수집하고 데이터가 무엇을 의미하는지를 한눈에 알 수 있도록 표현하는 것

✓ 미지의 결과를 추정, 예측하는 것

통계?

한정된 정보를 바탕으로, 복잡한 사회에서 무엇이 일어나고 있는지를 알기 쉽게 나타내고, 그로부터 어떤 일이 일어날지를 확률적으로 추계하는 수학

2) 통계의 필요성



A 방법 주장 :

B보다 2%나 차이가 나니까 우수하다.

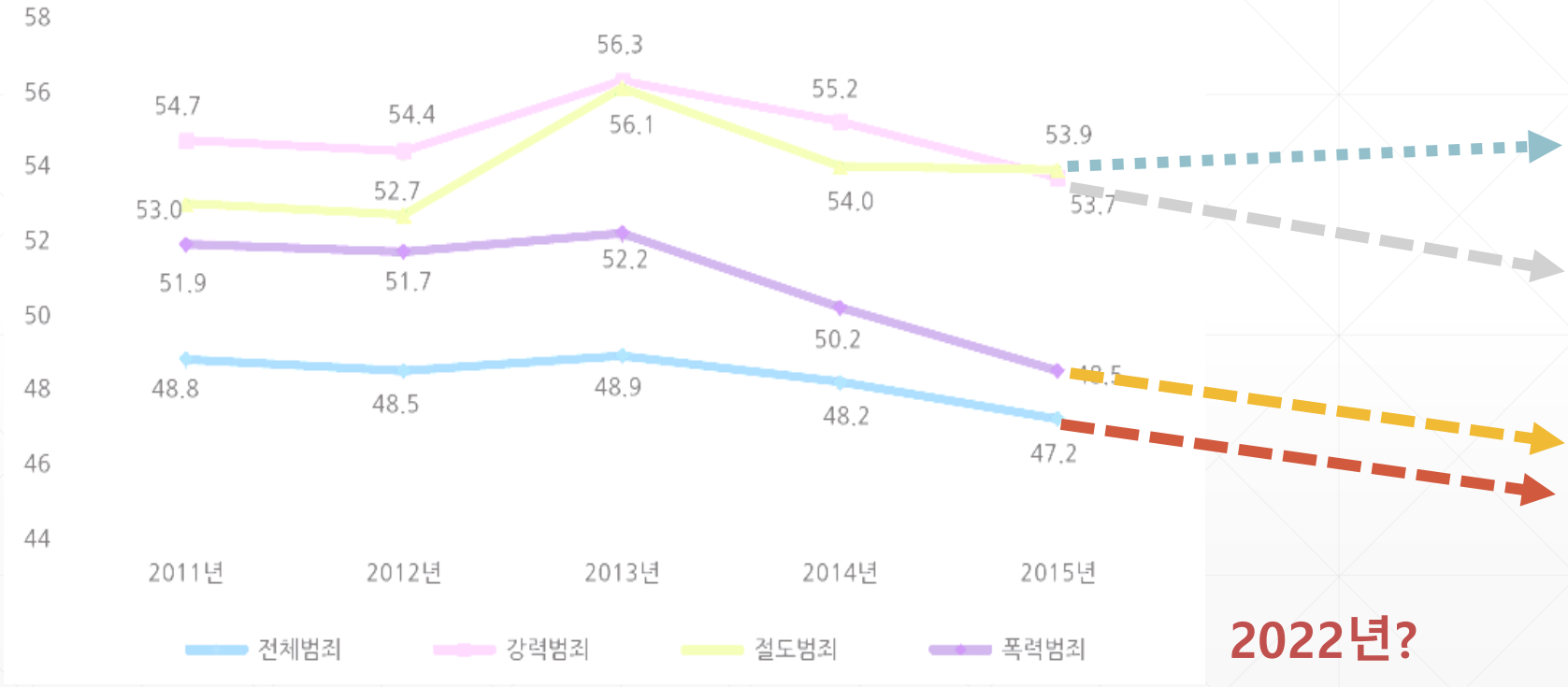
B 방법 주장:

B보다 2% 밖에 차이가 나지 않으니까 비슷하다.

집단 비교

2) 통계의 필요성

2022년도 범죄 발생 비율은?



2022년?

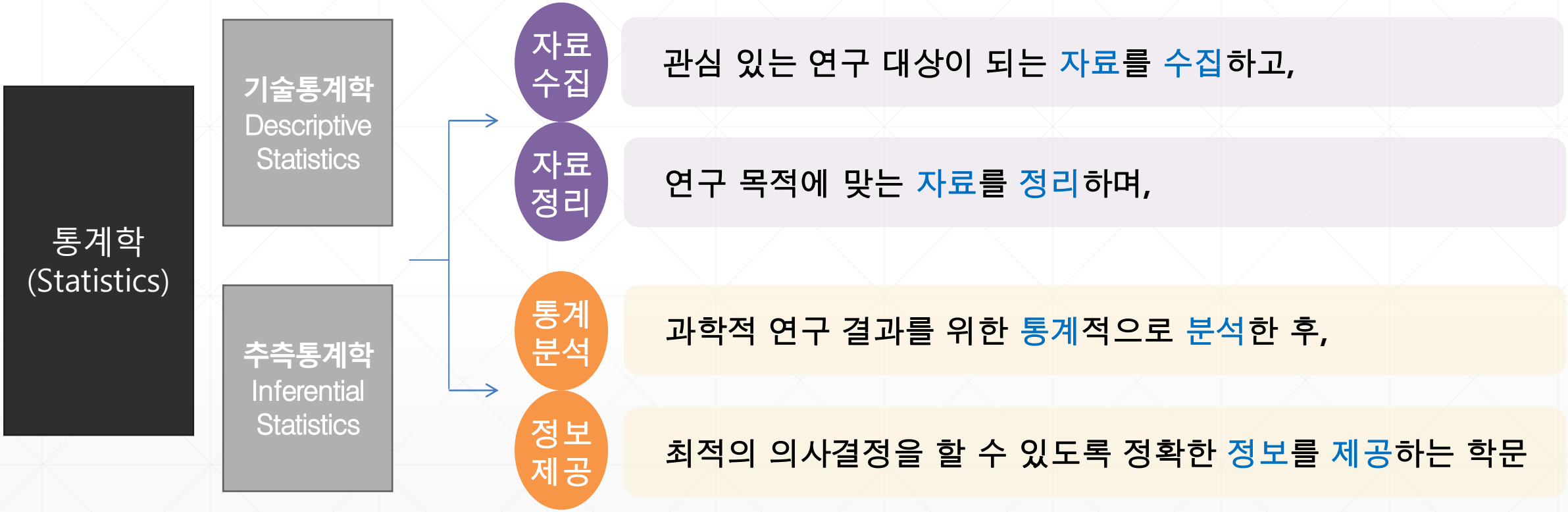
범죄율 예측

2) 통계의 필요성

https://www.etoday.co.kr/news/view/1209122

분류분석

3) 통계학(Statistics)



4) 기술통계학

기술통계학
(Descriptive Statistics)

수집한 자료의 특성을 적절히 묘사하기 위하여 자료를 정리, 요약, 계산, 기술하는 방법과 관련된 통계학

- 수집된 자료를 그래프, 표, 그림 또는 몇 개의 수치로 정리, 요약
- 자료의 전반적인 특징을 파악하기 위한 분석 방법을 다루는 분야

4) 기술통계학

✔ 자료의 요약 – 대표값

- 평균(Mean)
- 중위수(Median)
- 최빈값(Mode)

✔ 자료의 요약 – 산포도

- 분산(Variance)
- 표준편차(Standard Deviation)
- 변동계수(Coefficient of Variation)
- 범위(Range)
- 사분위 범위(Inter Quantile Range)
- 다섯 숫자 요약(5 Number Summary)

✔ 자료의 요약 – 형태

- 왜도(Skewness)
- 첨도(Kurtosis)

4) 기술통계학

✔자료의 요약 – 대표값

평균 (Mean)

- 보편적으로 사용하는 대표값
- 이상치(outlier)에 민감하게 영향

대조군

5, 5, 5, 10, 25

실험군

10, 10, 10, 10, 10

- ➢ 모든 경우에 평균이 적당한 건 아님.
- ➢ 이상치가 존재할 때는 중위수가 더 효과적
- ➢ 절사평균: 일정비율만큼 이상치 부분을 제거한 후 평균 산출

4) 기술통계학

✔자료의 요약 – 대표값

중위수 (Median)

- 전체 관측치를 크기 순으로 나열했을 때, 중앙에 위치한 중앙값
- 실제 데이터에 의존하지 않으므로 이상치(outlier)에 영향을 덜 받음

대조군

5, 5, 5, 10, 25

평균: 10 / 중위수: 5

실험군

10, 10, 10, 10, 10

평균: 10 / 중위수: 10

최빈값 (Mode)

- 전체 관측치들 가운데 가장 빈도가 높은 값
- 명목척도로 측정된 자료에 대한 대표값

4) 기술통계학

✔ 자료의 요약 – 산포도

분산
(Variance)

표본편차
(Standard Deviation)

- 자료의 흐트러진 정도를 나타내는 척도
- 분산 = (표본편차)^2
- 분산과 표준편차가 작을수록 자료는 평균에 근접하게 분포
- 관측치들의 측정단위와 표준편차의 측정단위는 같음
- 분산과 표준편차는 0보다 크거나 같음

변동계수
(Coefficient of Variation)

- 자료의 측정단위에 의존하지 않는 상대적인 산포의 측도
- 서로 측정단위가 틀린 여러 개의 자료의 산포를 비교할 때 사용
- 변동계수 = 표준편차/평균

4) 기술통계학

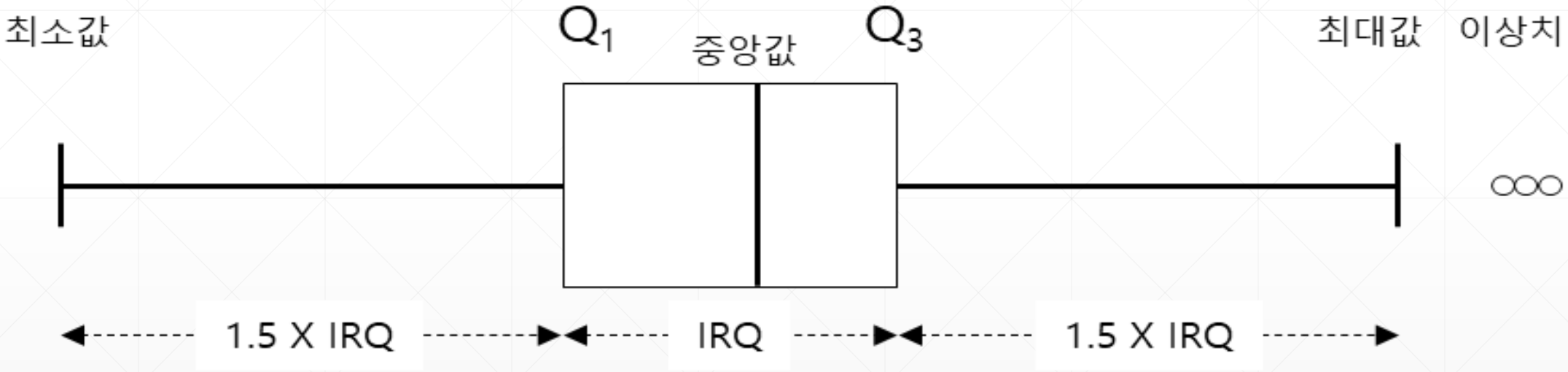
✔ 자료의 요약 – 산포도

범위 (Range)	• 범위 = 최대값–최소값 • 자료의 흐트러짐 정도를 나타내는 가장 간단한 측도 • 자료 중 이상치(outlier)에 민감
사분위 범위 (Inter Quantile Range)	• 사분위 범위 = 3사분위수–1사분위수 • 범위보다 이상치(outlier)의 영향을 덜 받음
다섯 숫자 요약 (5–Number Summary)	• 최소값, 1사분위수, 2사분위수, 3사분위수, 최대값

4) 기술통계학

✔자료의 요약 – 산포도

>>>> 상자그림(box plot)



Q1-1.5*(Q3-Q1)

Q3+1.5*(Q3-Q1)

IQR (InterQuartile Range, 사분위수 범위)

4) 기술통계학

✔ 자료의 요약 – 형태

왜도
(Skewness)

- 자료의 비대칭 정도를 나타내는 통계량

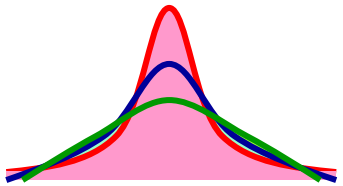
<0 (left-skewed)
오른쪽으로 치우쳐진 분포

=0
좌우대칭 분포

>0 (right-skewed)
왼쪽으로 치우쳐진 분포

첨도
(Kurtosis)

- 자료의 뾰족한 정도를 나타내는 통계량



=0 : 정규분포의 첨도
>0 : 정규분포보다 뾰족
<0 : 정규분포보다 더 평평

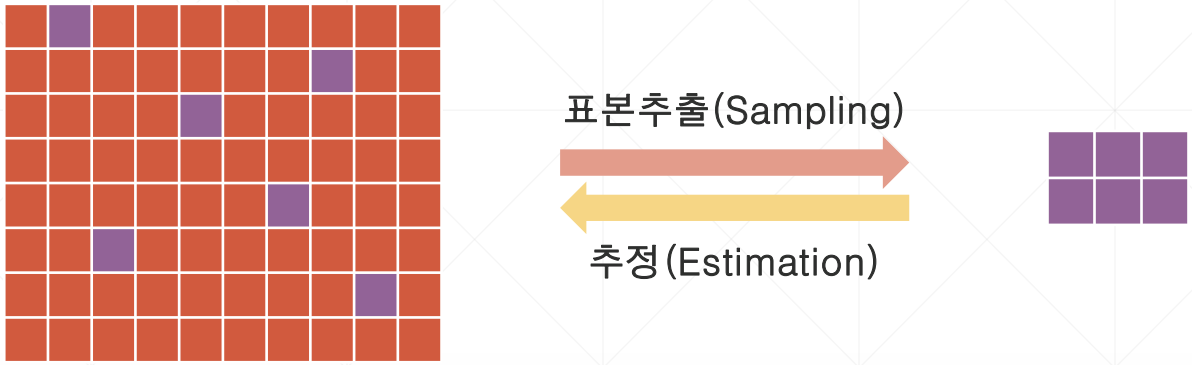
5) 추측통계학

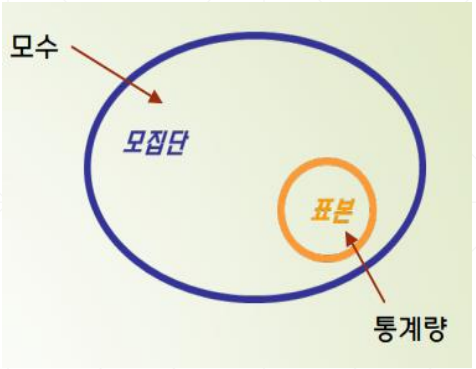
추측통계학
(Inferential Statistics)

자료에 내포되어 있는 정보를 분석하여 불확실한 사실에 대한 추론을 하는 분야이며, 모집단에서 추출한 표본을 분석하고 그 모집단에서 추출한 결과를 기초로 하여 모집단의 특성을 유추하고 일반화 하는 통계학

- 주어진 자료(표본)를 통해 불확실한 모집단의 특징을 추측, 추론
- 의사결정이나 미래예측을 하기 위한 분석 방법을 다루는 분야

5) 추측통계학





- **Population(모집단)**

-정보를 알고자 하는 대상들의 전체집합

- **Parameter(모수)**

-모집단의 특성치, 고정된 미지의 상수

-모평균, 모분산, 모비율…

- **Sample(표본)**

-모집단에 관한 정보를 담고 있는 자료의 일부분

- **Statistic(통계량)**

-표본의 특성치, 모수에 대한 통계적 추론을 위한 함수

-표본평균, 표본분산, 표본비율…

5) 추측통계학

조사 대상 선정과 표본 추출 단계에서 표집오차를 최소화하기 위해 모집단의 특성을 고려하고 무작위 표본 추출을 통해 적정 수 이상의 표본 수 확보

표집 오차

✓**정의:** 표본을 추출해 조사하기 때문에 발생되는 오차

✓**원인:** 무작위성과 동질성

✓**해결방법:** 편의 제거, 적절한 표본추출 방법 선택, 적절한 표본의 크기 확보

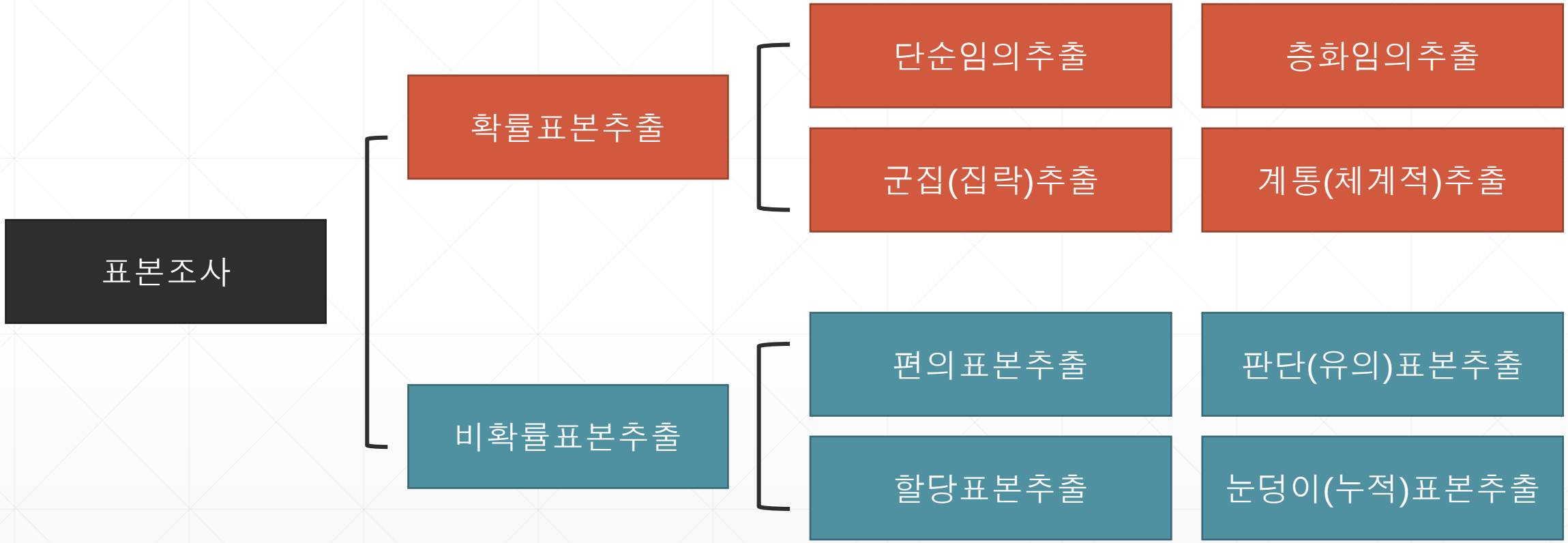
비표집오차

✓**정의:** 표본 추출하는 것과 관계 없는 오차

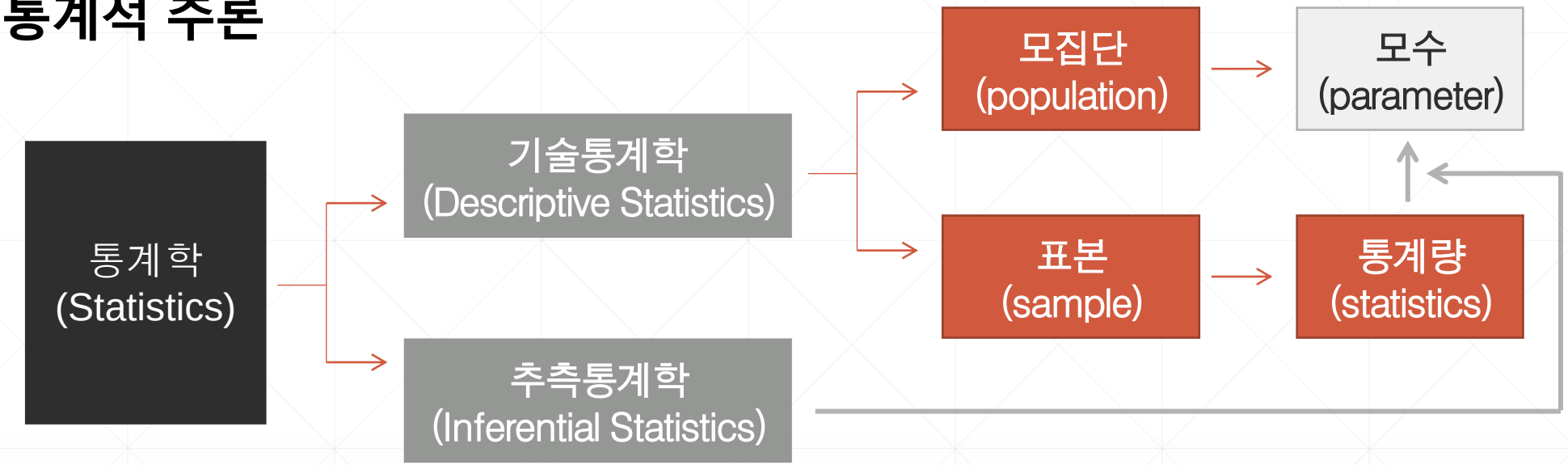
✓**원인:** 무응답오차, 응답오차, 처리오차

✓**해결방법:** 응답자의 관심 유발, 조사원의 철저한 교육, 정확한 통계 처리

5) 추측통계학


표본조사
확률표본추출
비확률표본추출
단순임의추출
층화임의추출
군집(집락)추출
계통(체계적)추출
편의표본추출
판단(유의)표본추출
할당표본추출
눈덩이(누적)표본추출

6) 통계적 추론



- 추정(estimation), 가설검정(test of hypothesis)

- **통계적 추론(statistical inference)**

: 표본으로부터의 정보를 이용하여 모집단에 대한 추측 또는 의사결정을 하는 과정

: 접근 방법에 따라 모수적 방법, 비모수적 방법, 베이즈적 방법

6) 통계적 추론

귀무가설(Null hypothesis, H_0)

- 영가설
- 의미적으로 차이를 나타내지 않는 경우의 가설

대립가설(Alternative hypothesis, H_1)

- 귀무가설과 반대되는 가설
- 연구자가 연구를 통해 입증하고자 하는 가설

1종 오류(Type I error)

- 귀무가설이 사실일 때, 귀무가설을 기각하는 오류

2종 오류(Type II error)

- 귀무가설이 거짓일 때, 귀무가설을 기각하지 않는 오류

유의수준(Significance level)

- type I error의 허용 한계

유의확률(significance probability)

- 귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률

검정력(Power of Test)

- 귀무가설이 사실이 아닐 때, 귀무가설을 기각할 확률$(1-\beta)$

신뢰구간(Confidence Interval)

- 모수가 실제로 포함될 것으로 예측되는 구간

7) 가설검정(Hypothesis test)

귀무가설(Null hypothesis, H_0)

- 영가설
- 의미적으로 차이를 나타내지 않는 경우의 가설

대립가설(Alternative hypothesis, H_1)

- 귀무가설과 반대되는 가설
- 연구자가 연구를 통해 입증하고자 하는 가설

7) 가설검정(Hypothesis test)

이 사람은 왕자이다.

반박 이론:

1) 왕자는 남자인데 여자?

2) 왕자는 왕실에 사는 사람이니 좋은 음식을 먹을 텐데, 너무 말라서 평민일지도?

3) 왕자라고 하기에는 나이가 너무…

4) 사람이 아니라 외계인일지도?

…

7) 가설검정(Hypothesis test)

비만이면 암에 걸릴 가능성이 높아진다.

반박 이론:

1) 비만, 암이라는 기준은?

2) 비만이 아니라 성별, 연령 때문 아님?

3) 비만이 바로 암에 영향을 주는 것이 아니라 비만 -> 만성질환 -> 암 아님?

4) 맞게 분석한 거임?

…

7) 가설검정(Hypothesis test)

- H_0 : 공주이다
- H_1 : 왕자이다

➢ 검정 결과 : P = 0.874

☞ P>0.05 : 대립가설 기각

☞ 따라서, “공주이다”?

☞ 대립가설(왕자이다)만 기각

☞ 귀무가설(공주이다)이 맞다는 것은 아니다

7) 가설검정(Hypothesis test)

- H_0 : 공주이다.
- H_1 : 왕자이다.

- H_0 : A=공주
- H_1 : A≠공주

- H_0 : A=왕자
- H_1 : A≠왕자

7) 가설검정(Hypothesis test)

- 남학생, 여학생 몸무게 차이 비교
- 통상적으로 몸무게 차이가 3kg 보다 작으면 같다고 가정

➢ 몸무게 차이 비교

- H_0 : 남학생 = 여학생
- H_1 : 남학생 ≠ 여학생

양측검정

- H_0 : 남학생 = 여학생
- H_1 : 남학생 >(or <) 여학생

단측검정

➢ 몸무게 차이가 나지 않음을 비교

- H_0 : 남학생 ≠ 여학생
- H_1 : 남학생 = 여학생

- H_0 : 남학생 = 여학생
- H_1 : 남학생 ≠ 여학생

P >0.05 (?)

7) 가설검정(Hypothesis test)

- 남학생, 여학생 몸무게 차이 비교
- 통상적으로 몸무게 차이가 3kg 보다 작으면 같다고 가정

- H_0 : |남학생 몸무게 – 여학생 몸무게| = 3kg
- H_1 : |남학생 몸무게 – 여학생 몸무게| < 3kg

단측검정

비열등성 시험 (Non-Inferiority Trial)	동등성 시험 (Equivalence Trial)
➢ 기존약에 비해 신약의 효과가 떨어지지 않는다.	➢ 기존약과 신약의 효과는 같다.

8) 가설오류

1종 오류(Type I error)

- 귀무가설이 사실일 때, 귀무가설을 기각하는 오류

2종 오류(Type II error)

- 귀무가설이 거짓일 때, 귀무가설을 기각하지 않는 오류

유의수준(Significance level)

- type I error의 허용 한계

유의확률(significance probability)

- 귀무가설이 맞다고 가정할 때 얻은 결과보다 극단적인 결과가 실제로 관측될 확률

검정력(Power of Test)

- 귀무가설이 사실이 아닐 때, 귀무가설을 기각할 확률$(1-\beta)$

신뢰구간(Confidence Interval)

- 모수가 실제로 포함될 것으로 예측되는 구간

2. 통계분석 방법론

1) 변수

질적자료(Qualitative Data)

- 범주로 구성
- **명목척도** : 1. 귀하의 성별은? ①남자 ②여자
  2. 귀하의 거주지는? ①서울 ②부산 ③대구 ....
- **서열척도** : 측정대상간의 서열관계 존재
  ex) 성적(A, B, C…), 순위(1위, 2위, 3위…)

양적자료(Quantitative Data)

- 연속된 수로 구성
- **등간척도** : 간격 동일, 연산 가능, 숫자 **0의 의미 없음**
  ex)온도…(100℃가 반드시 50℃보다 2배 높은 온도라고 말할 수는 없다.)
- **비율척도** : 연산 가능, 숫자 **0의 의미가 있음**
  ex)인구수, 키, 몸무게, 시간…

2) 변인과의 관계

Association(연관관계)	둘 이상의 변수 간의 관계성
Correlation(상관관계)	둘 이상의 변수 간의 (선형) 관계성
Causation(인과관계)	둘 이상의 변수 간의 (원인-결과) 관계성
Interaction(교호관계)	둘 이상의 변수 간의 (한 변수의 효과가 다른 변수의 효과에 영향을 주는) 관계성

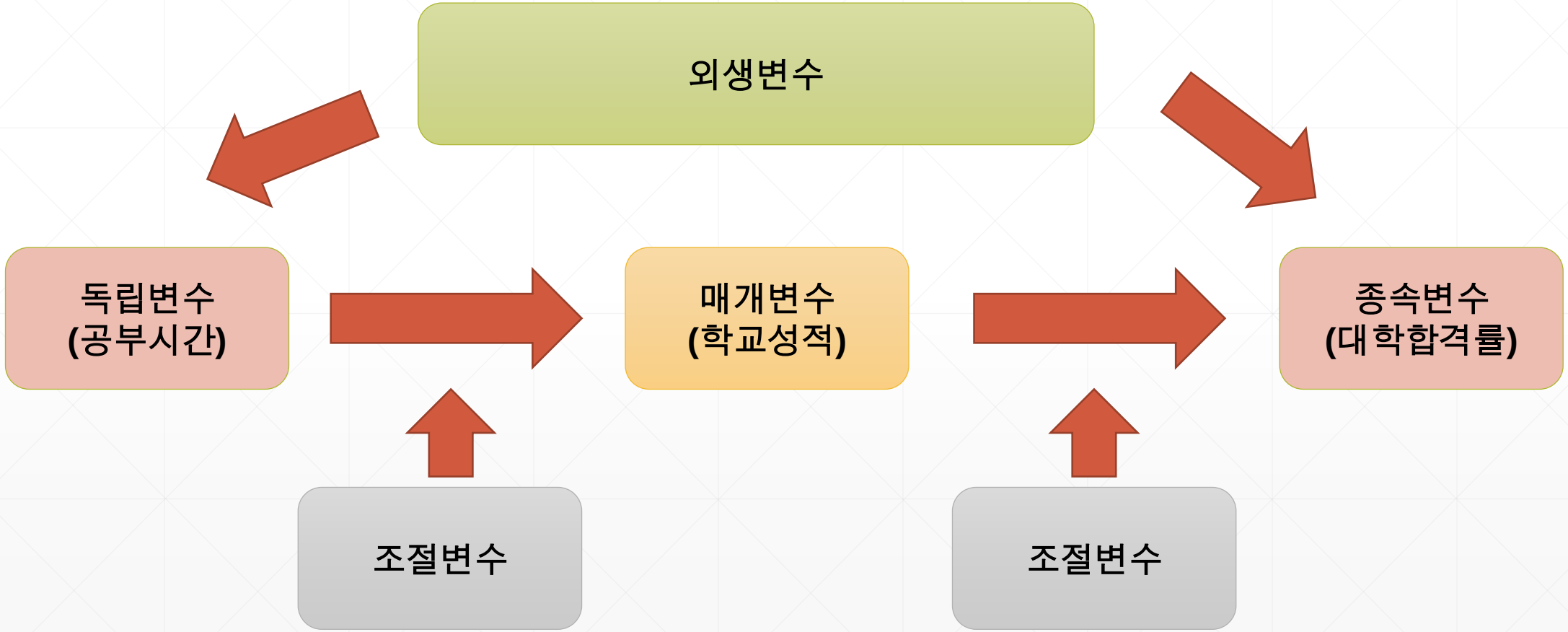
2) 변인과의 관계

공부시간은 대학 합격률에 영향을 준다.

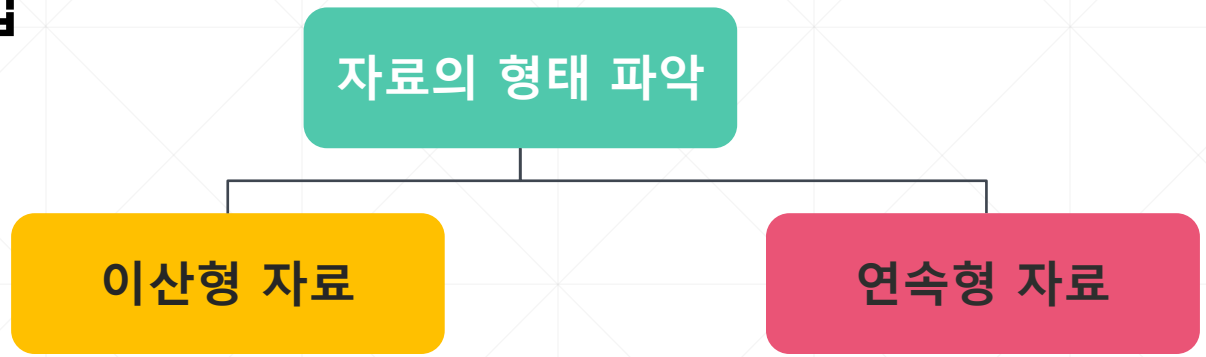
독립변수(independent variable)	종속변인에 영향을 주는 변인 ex) 공부시간
종속변인(dependent variable)	독립변인에 영향을 받는 변인 ex) 대학 합격률
매개변인(mediating variable)	독립변인에게 영향을 받고, 종속변인에 영향을 주는 변인 ex) 학교 성적
조절변인(moderating variable)	독립변인이 종속변인에 미치는 영향의 강도에 영향을 주는 변인 ex) 지능
통제변인(control variable)	연구에 영향을 주지 않도록 신경 써야 하는 요인 ex) 성별, 연령
외생변인(Extraneous variable)	독립변인과 종속변인 모두 영향을 미치는 숨은 변인 ex) 학원수가 늘어나면 대학 합격률이 증가한다.

2) 변인과의 관계

공부시간은 대학 합격률에 영향을 준다.

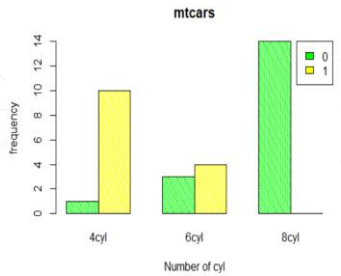


3) 통계분석 방법

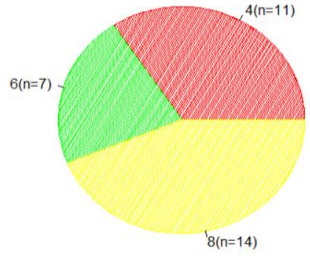


3) 통계분석 방법

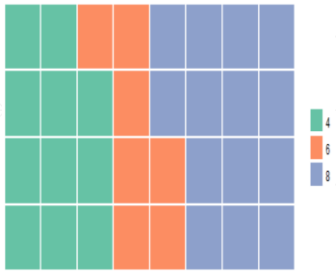
이산형 자료



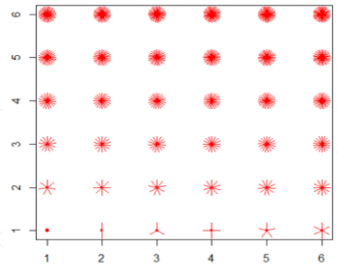
막대 그래프
(bar plot)



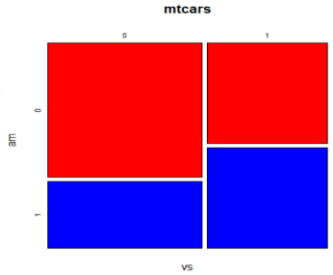
파이 그래프
(pie plot)



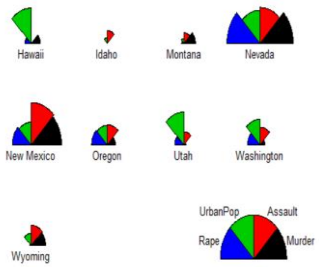
와플 그래프
(waffle plot)



해바라기 그래프
(sunflower plot)

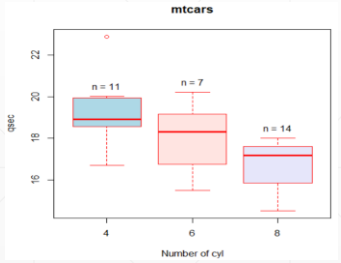


모자이크 그래프
(mosaic plot)

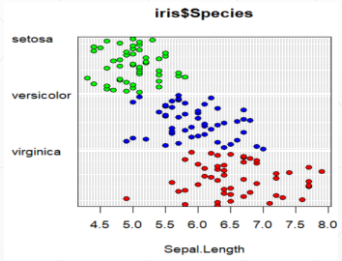


별 그래프
(stars plot)

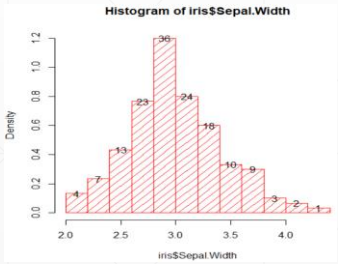
연속형 자료



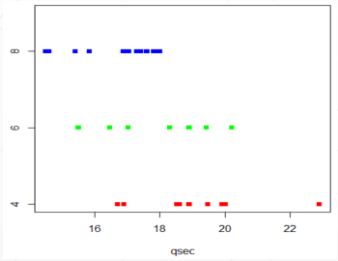
상자 그래프
(box plot)



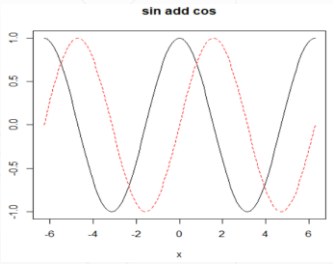
점 그래프
(dot plot)



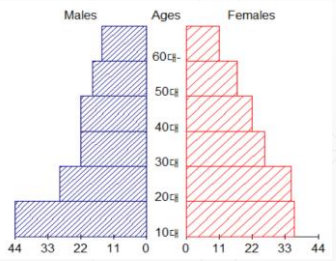
히스토그램
(histogram)



띠 그래프
(pie plot)

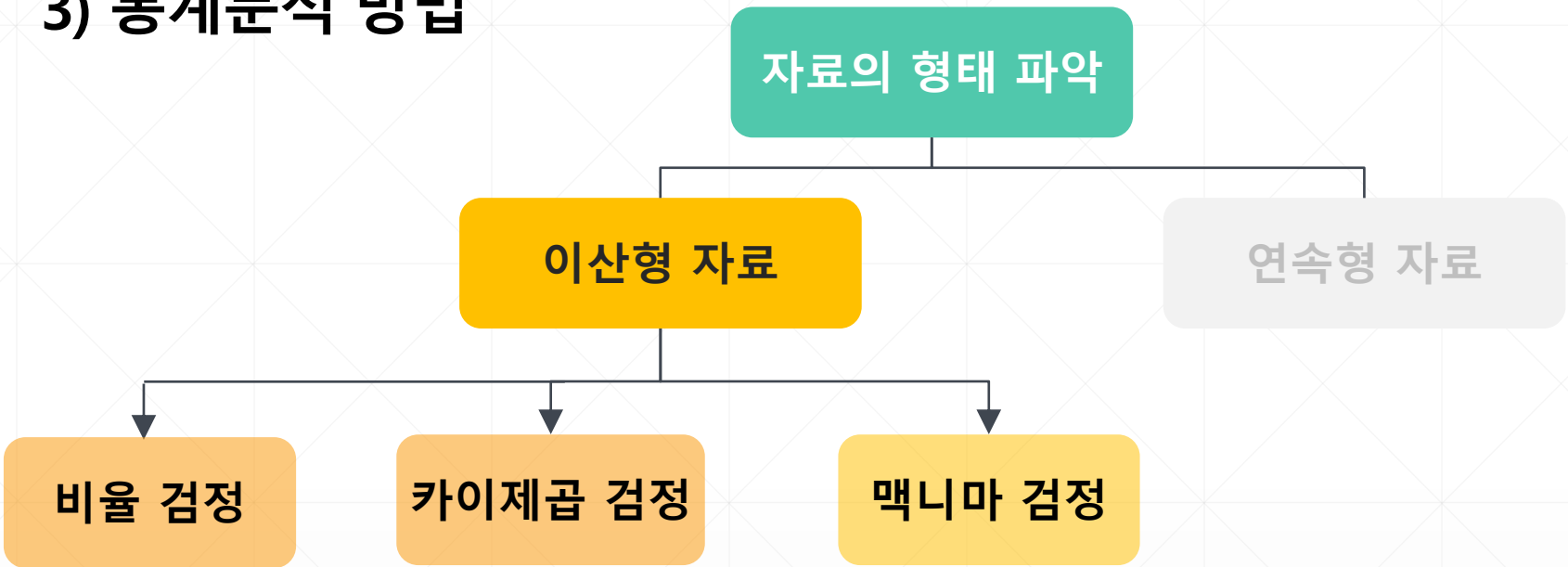


커브 그래프
(curve plot)



인구피라미드

3) 통계분석 방법



모집단이 1개

모집단이 2개 이상

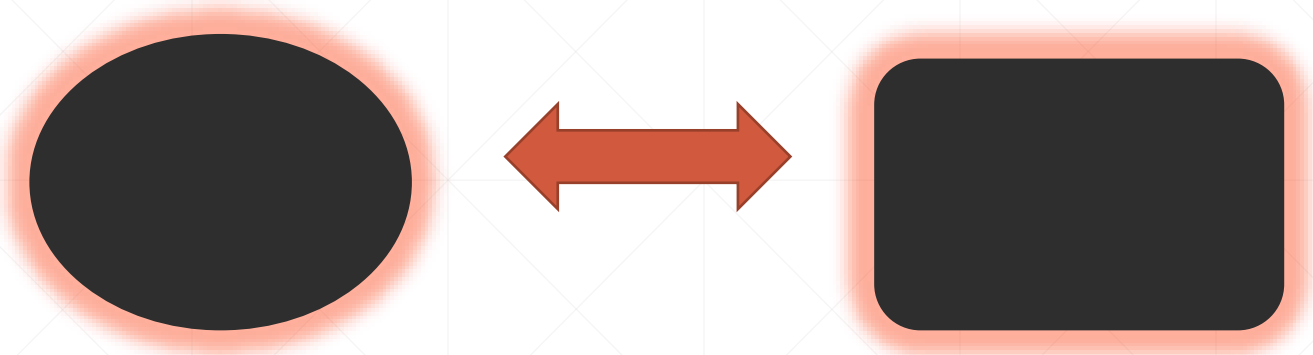
독립적인 자료끼리의
교차 분석

짝지은 자료끼리의
교차 분석

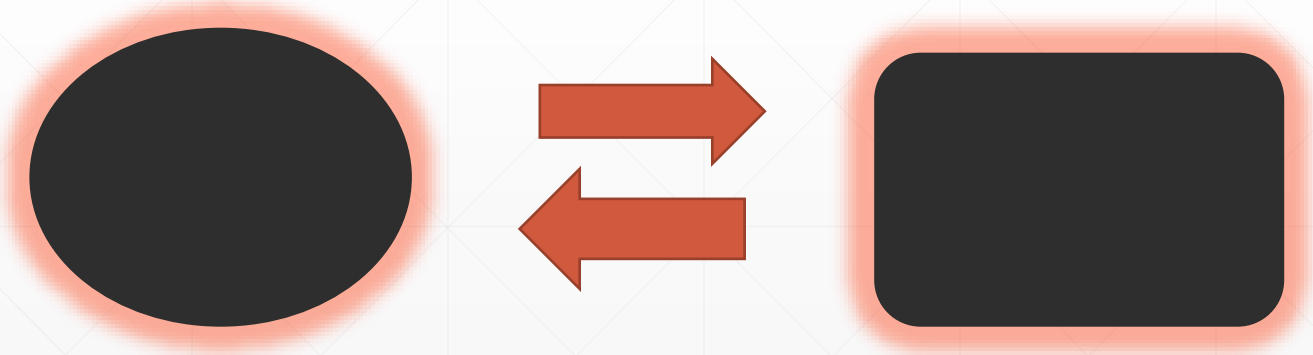
3) 통계분석 방법


자료의 형태 파악
이산형 자료
연속형 자료
비율 검정
카이제곱 검정
맥니마 검정
정규성 검정
모수적 방법
비모수적 방법
일표본 T 검정
독립표본 T 검정
대응표본 T 검정
일원배치 분산분석
Signed Rank Test
Wilcoxon’s Rank Sum Test
차이값을 계산하여 Signed Rank Test
Kruskal-Wallis test
사후검정

3) 통계분석 방법



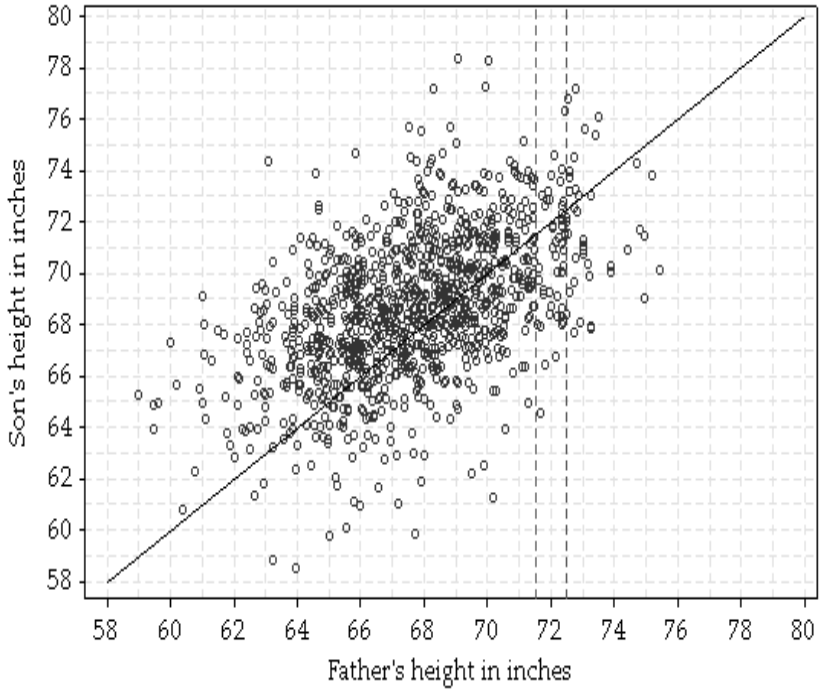
상관분석 : 방향성에 대한 고려 없음. 단순히 선형 상관성이 있음을 확인



회귀분석 : 원인, 결과에 대한 방향 고려.

3) 통계분석 방법

회귀분석(Regression Analysis)

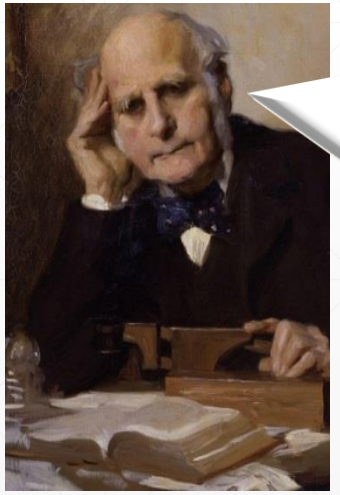


The figure is based on a graph from Statistics by Freedman, Pisani, Purves and Adhikari(http://www.scc.ms.unimelb.edu.au/whatisstatistics/faso.html)

- 왜 회귀(回歸)인가 ?

  - 아버지가 크면 아들도 큰가?

  - 세대가 갈수록 키가 큰 집단과 작은 집단만 생기나?



아버지의 키가 큰 경우
아들의 키는 평균키로 하향하고,
아버지의 키가 작은 경우 아들의 키는 역시 평균키로 상향한다.
-Francis Galton

https://en.wikipedia.org/wiki/Francis_Galton

3) 통계분석 방법

회귀분석(Regression Analysis)

- 변수들 사이의 함수적 관계를 탐색하는 방법
- **독립(independent) 변수**: 실험에 영향을 줄 것이라고 예상되는 변수, 설명 변수(explanatory variable)
- **종속(dependent) 변수**: 독립변수에 영향을 받아 결과적으로 변화될 것이라고 예상되는 변수, 반응 변수(response variable)
- 인과관계 : Y(종속변수) ← X(독립변수)

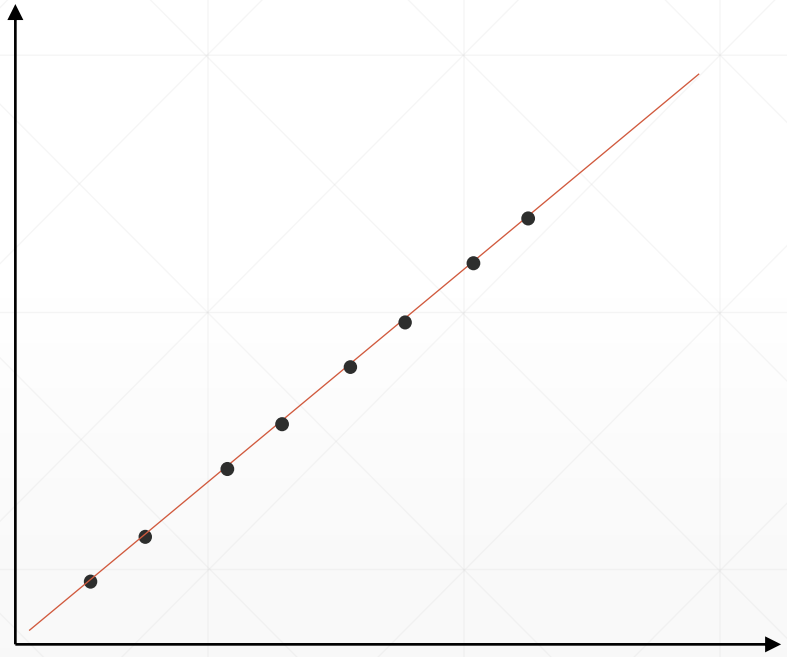
$$Y = B_0 + B_1 X$$

: 회귀(모형)식

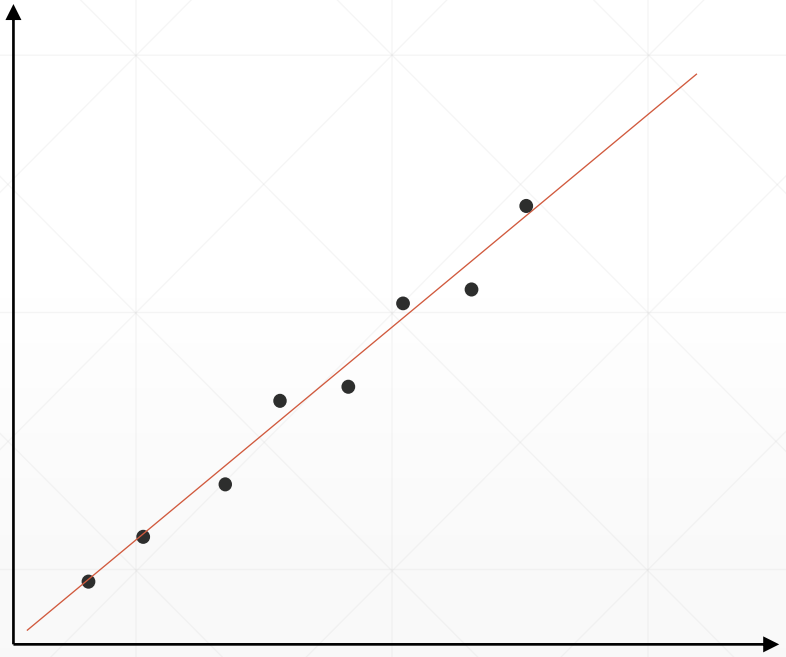
회귀계수

3) 통계분석 방법

회귀분석(Regression Analysis)



$$Y = B_0 + B_1X$$



$$Y = B_0 + B_1X + \varepsilon$$

3) 통계분석 방법

회귀분석 조건

$$Y = B_0 + B_1X + \varepsilon$$

i) 오차항 ε_i의 기대값은 0이다. ($E(\varepsilon_i)= 0$)

ii) 오차항 ε_i는 모두 동일한 분산을 갖는다.($Var(\varepsilon_i)= \sigma^2$)

iii) 오차항 ε_i는 정규분포를 이루며, 서로 독립적이다.($Cov(\varepsilon_i , \varepsilon_j)= 0(i \neq j)$)

-독립변수와 종속변수간의 선형성(변화가 일정)

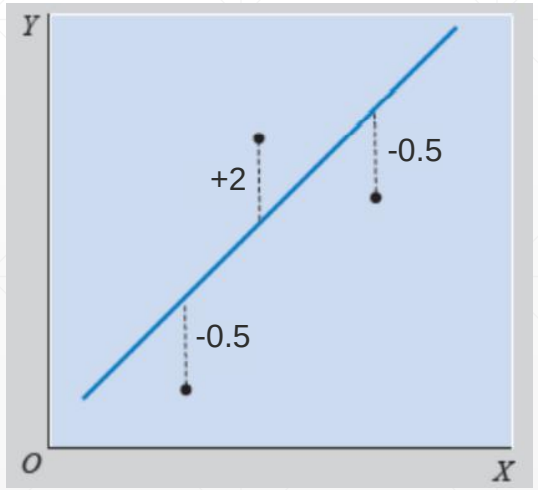
-오차항의 독립성, 등분산성, 정규성 만족

3) 통계분석 방법

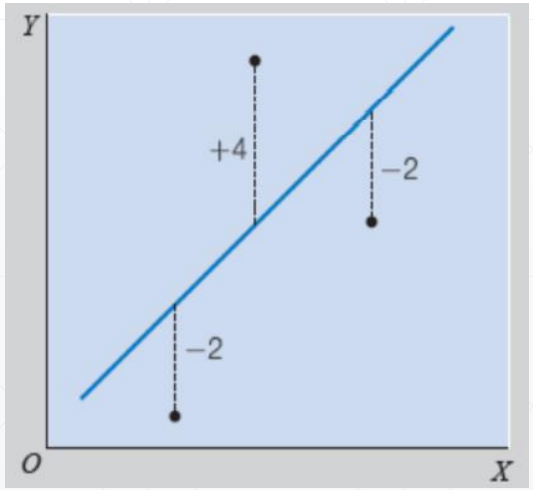
회귀선 추정

측정치를 가장 잘 나타낼 수 있는 가상의 선을 긋고 실제 관측치와 모형에 의한 예측치 간 거리의 잔차(residual)을 극소화하는 방법을 이용하여 β0, β1 의 계수를 구하는 방법

1) min(sum($Y_i - \hat{Y}_i$))



잔차=+1



잔차=0

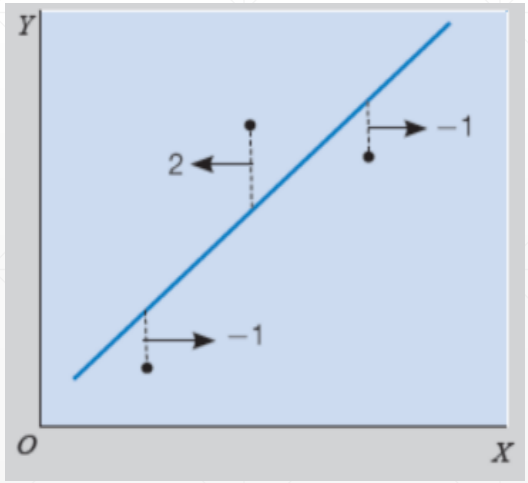
부호 고려 못함

3) 통계분석 방법

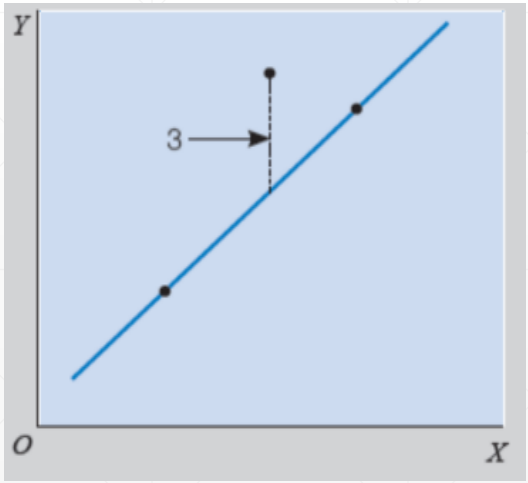
회귀선 추정

측정치를 가장 잘 나타낼 수 있는 가상의 선을 긋고 실제 관측치와 모형에 의한 예측치 간 거리의 잔차(residual)을 극소화하는 방법을 이용하여 β0, β1 의 계수를 구하는 방법

2) MAD(minimum absolute deviation) : min(sum($|Y_i - \hat{Y}_i|$))



MAD=4



MAD=3

3) 통계분석 방법

회귀선 추정

3) 최소제곱법(Ordinary Least Squares : OLS): min(sum(($Y_i - \hat{Y}_i$)^2))

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ i=1, 2, ... , n

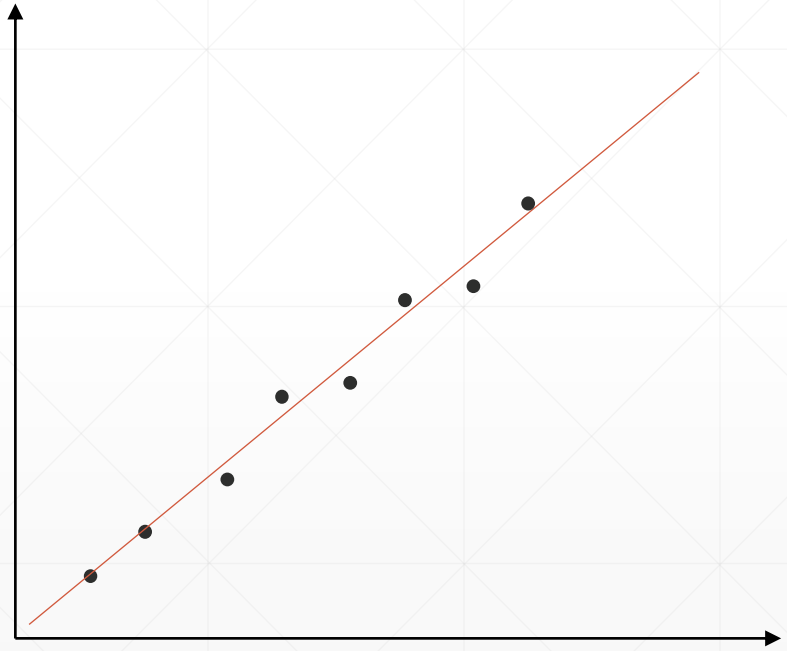
$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2$$: 잔차 제곱합

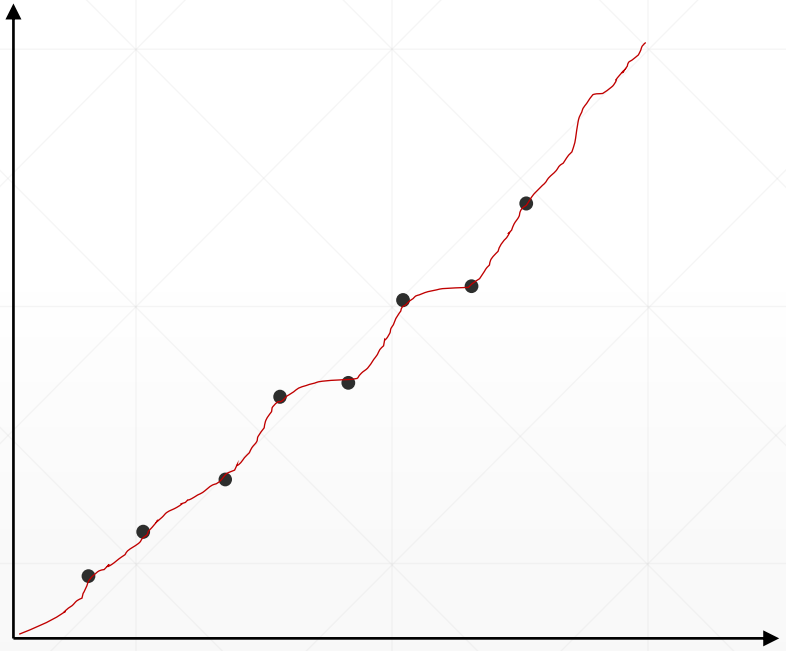
잔차제곱합이 최소가 되는 최소제곱추정량(LSE, least squares estimator)인 $\hat{\beta}_0$, $\hat{\beta}_1$ 을 구함.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

3) 통계분석 방법

회귀선 추정





어느 것이 좋은 회귀 추정선일까?

3) 통계분석 방법

회귀분석 해석

$$Y = B_0 + B_1X + \varepsilon\ ,\ \varepsilon \sim N(0, \sigma^2)$$

독립변수 X가 한 단계 늘어남에 따라 종속변수 Y는 B_1만큼 늘어나게 된다.

3) 통계분석 방법

회귀분석 해석

$$Y = B_0 + B_1X + \varepsilon\ ,\ \varepsilon \sim N(0, \sigma^2)$$

Y가 연속형이 아니라 범주형이라면?

1) Y = 0 or 1 범주형

2) P(Y=0), P(Y=1) 로 표현 가능

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = B_0 + B_1X$$

3) 통계분석 방법

회귀분석 해석

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = B_0 + B_1X + \varepsilon\ ,\ \varepsilon \sim N(0, \sigma^2)$$

$\exp(B_1)$: 오즈비(odds ratio)

독립변수 X가 한 단계 늘어남에 따라 종속변수 Y가 1이 확률이 Y가 0이 될 확률에 비해 $\exp(B_1)$배만큼 늘어난다.

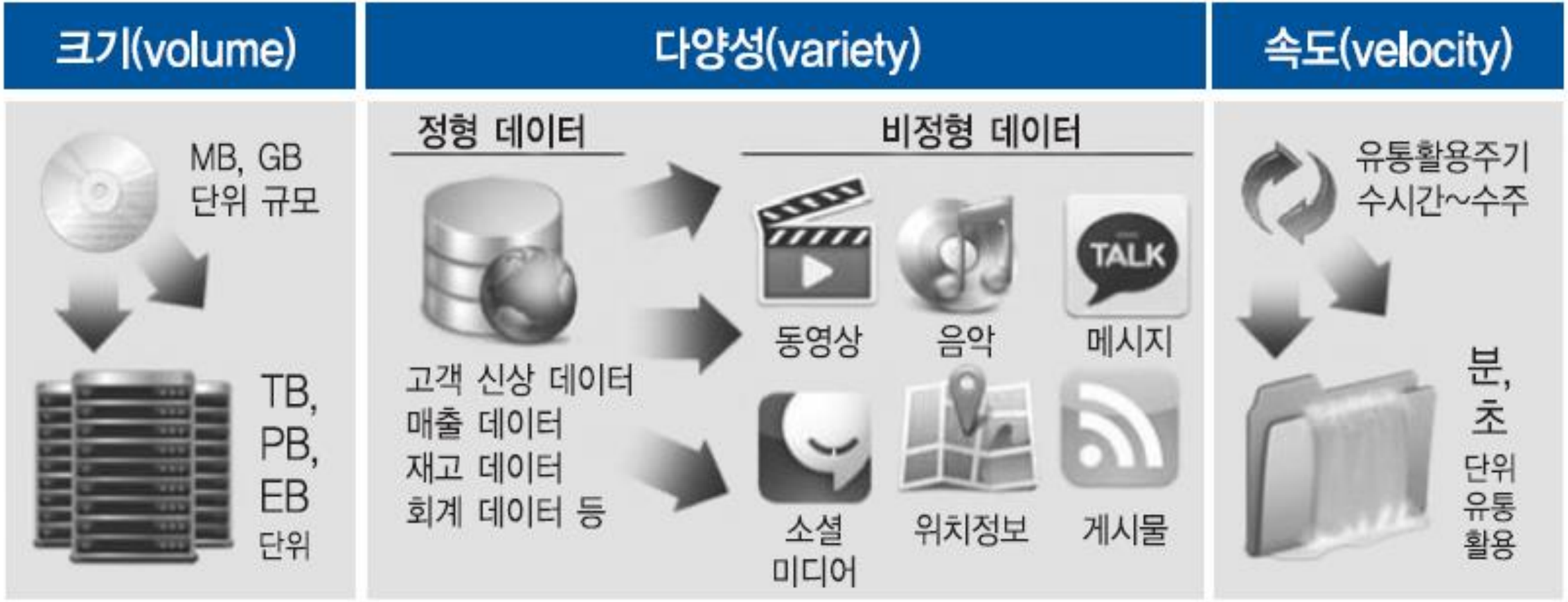
(=오즈비)

3) 통계분석 방법

	종속변수	독립변수	분석방법
$Y=\beta_0+\beta_1X$			
$Y=\beta_0+\beta_1X_1+\ldots+\beta_nX_n+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$	1개, 연속형 자료	연속형 자료	회귀분석
	1개, 연속형 자료	범주형 자료	분산분석
	1개, 연속형 자료	범주형 자료, 공변량	공분산분석
	1개, 범주형 자료	연속형, 범주형 자료	로지스틱 회귀분석
$Y_1+\ldots+Y_n=\beta_0+\beta_1X+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$	3개 이상(반복측정)	연속형, 범주형 자료	반복측정 분산분석
	사건, 중도절단, 기간	연속형, 범주형 자료	생존분석
$Y_{1i}=\beta_0+\beta_1X_i+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma_i^2)$	시간 자료	연속형, 범주형 자료	시계열 분석

3. 빅데이터 자료 분석 접근

1) 빅데이터란?

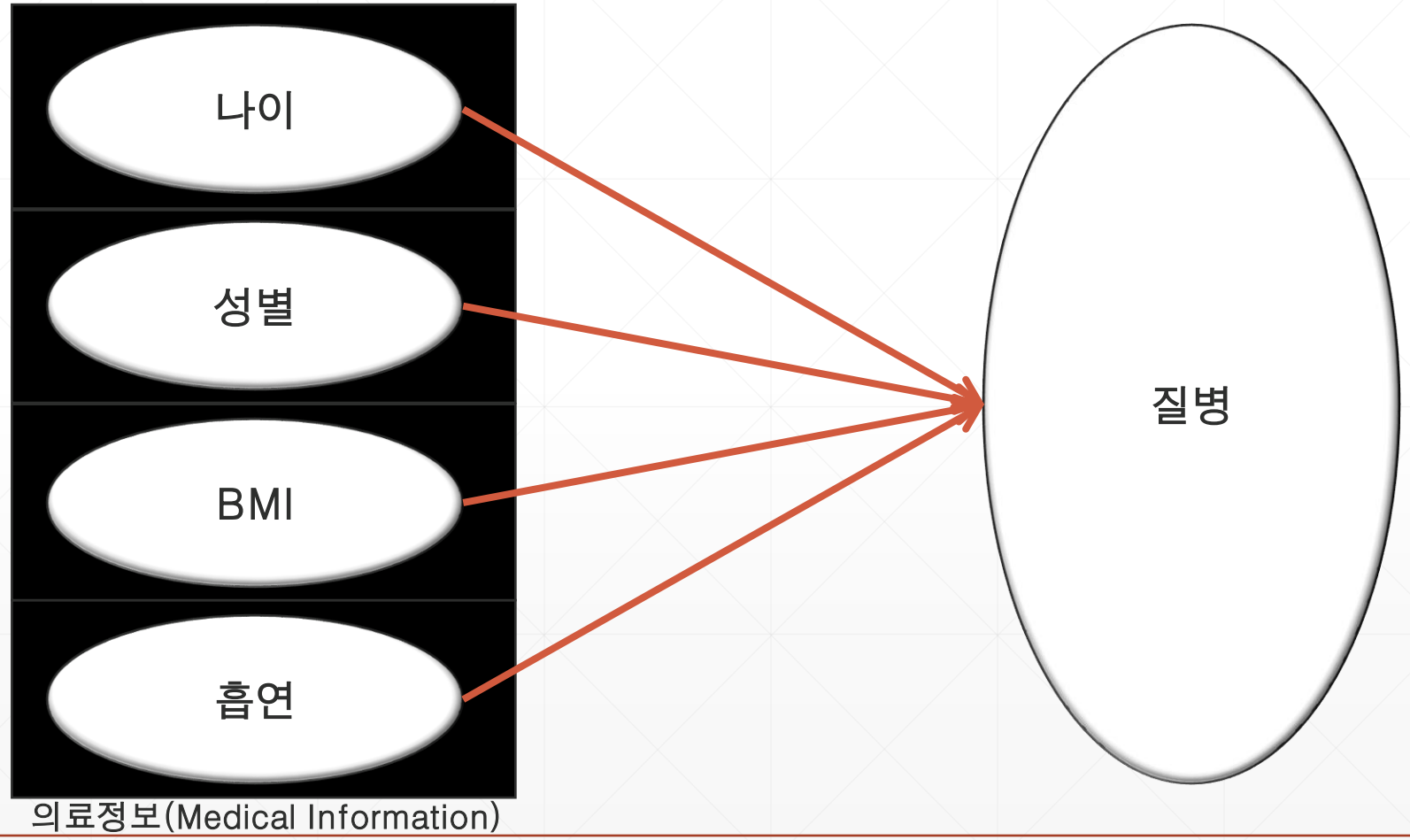


출처: 채승병 외, 빅데이터: 산업 지각변동의 진원, SERI CEO Information, 2012. 5

3. 빅데이터 자료 분석 접근

1) 빅데이터란?

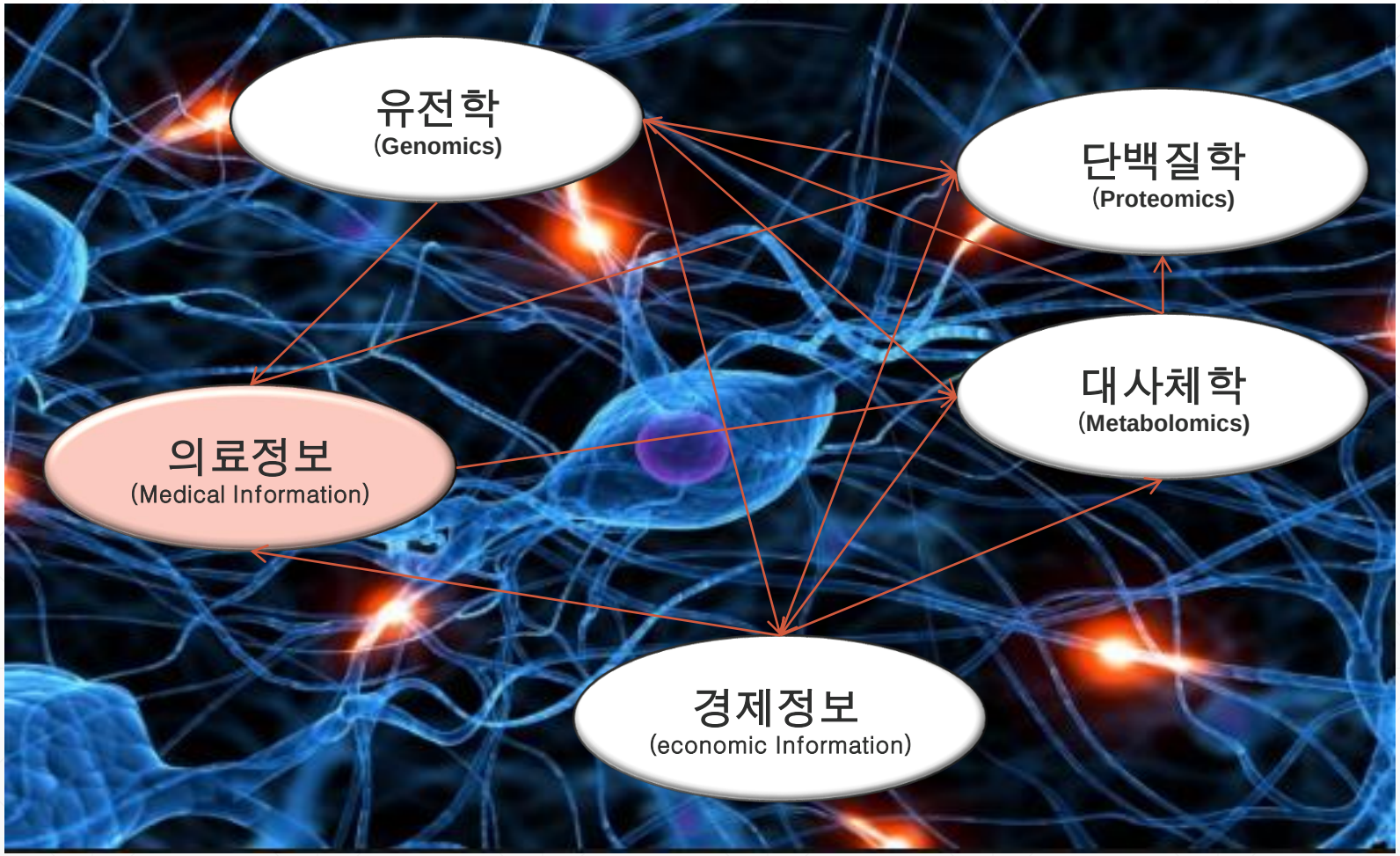
단일 분석



3. 빅데이터 자료 분석 접근

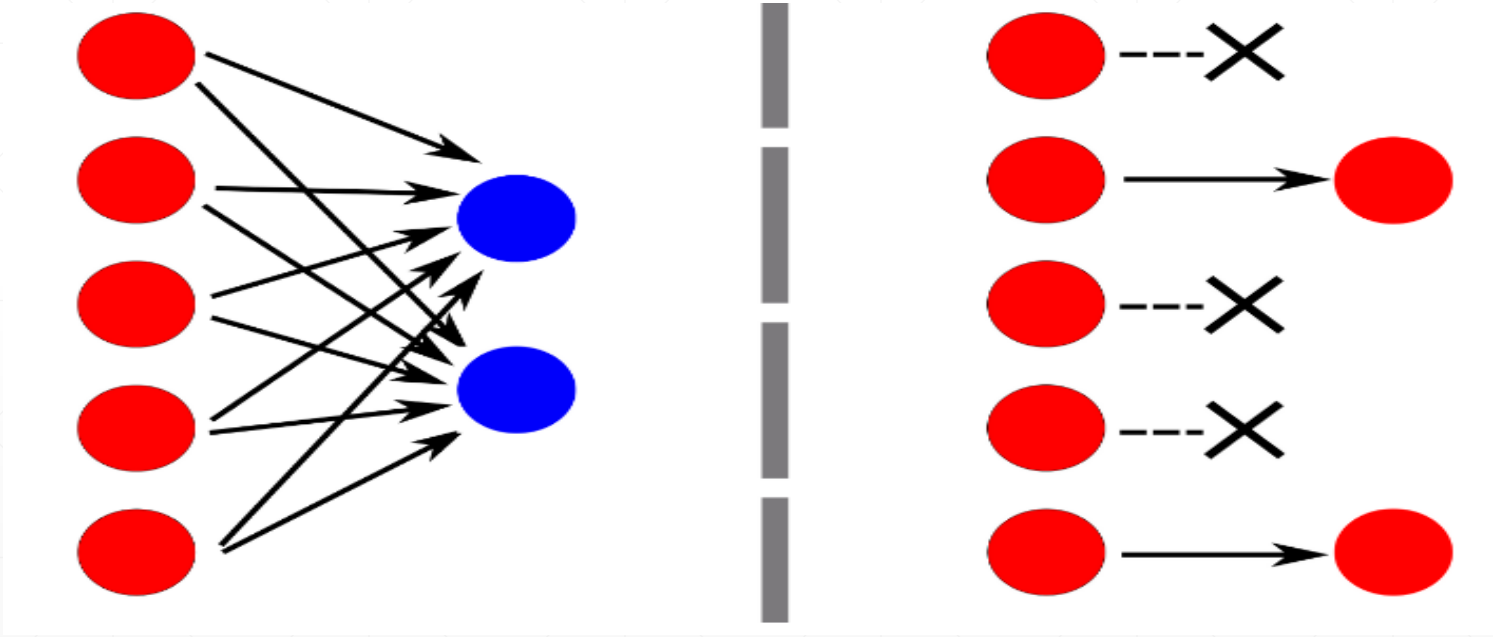
1) 빅데이터란?

네트워크식 통합 분석



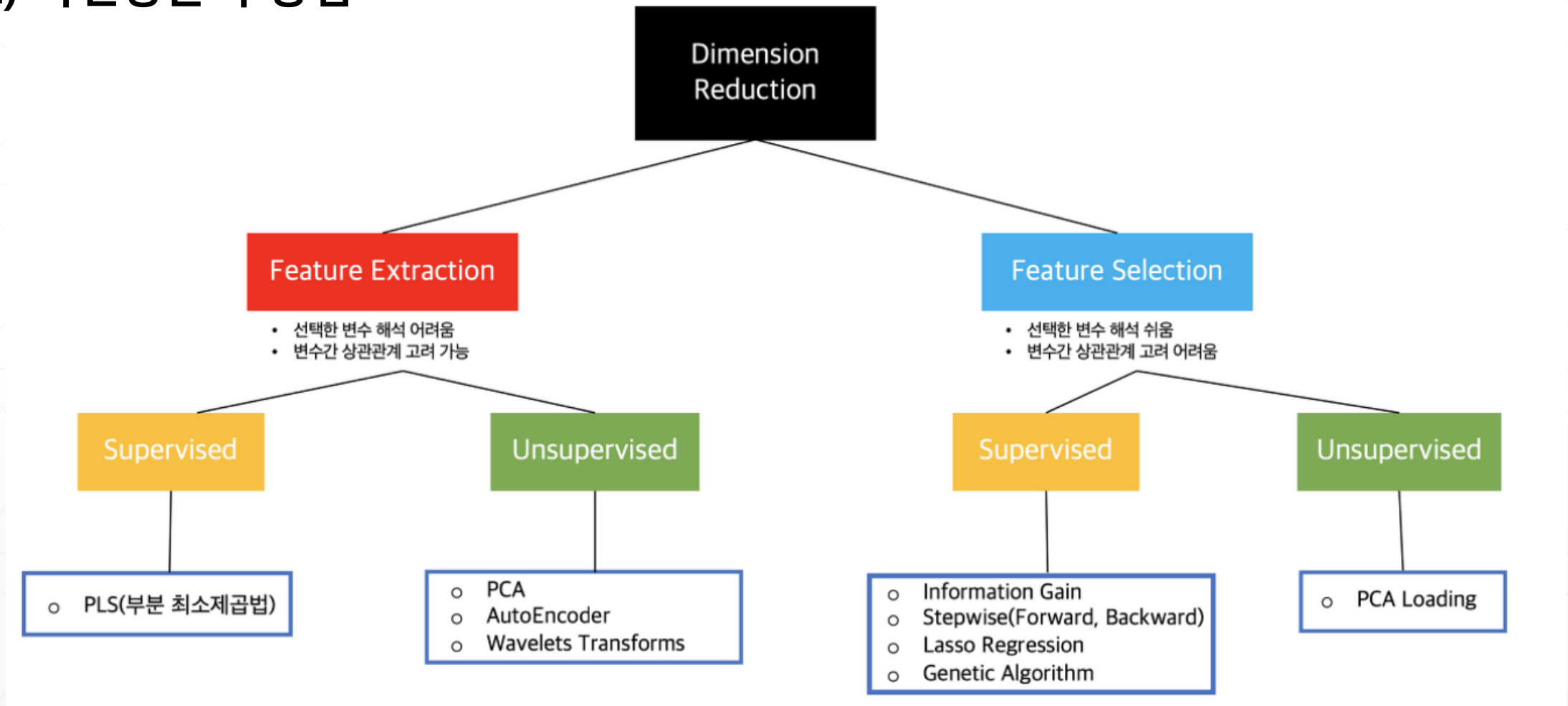
2) 다변량분석 방법

** 차원 축소

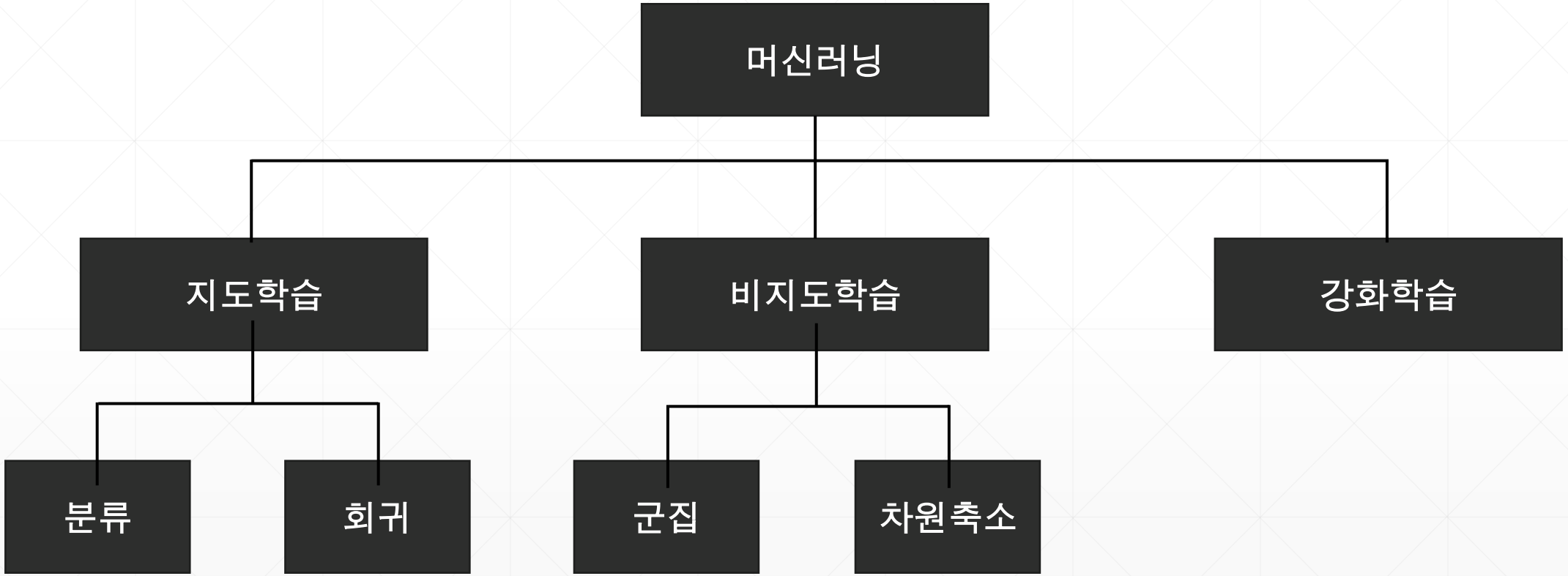


변수 축소 변수 선택

2) 다변량분석 방법

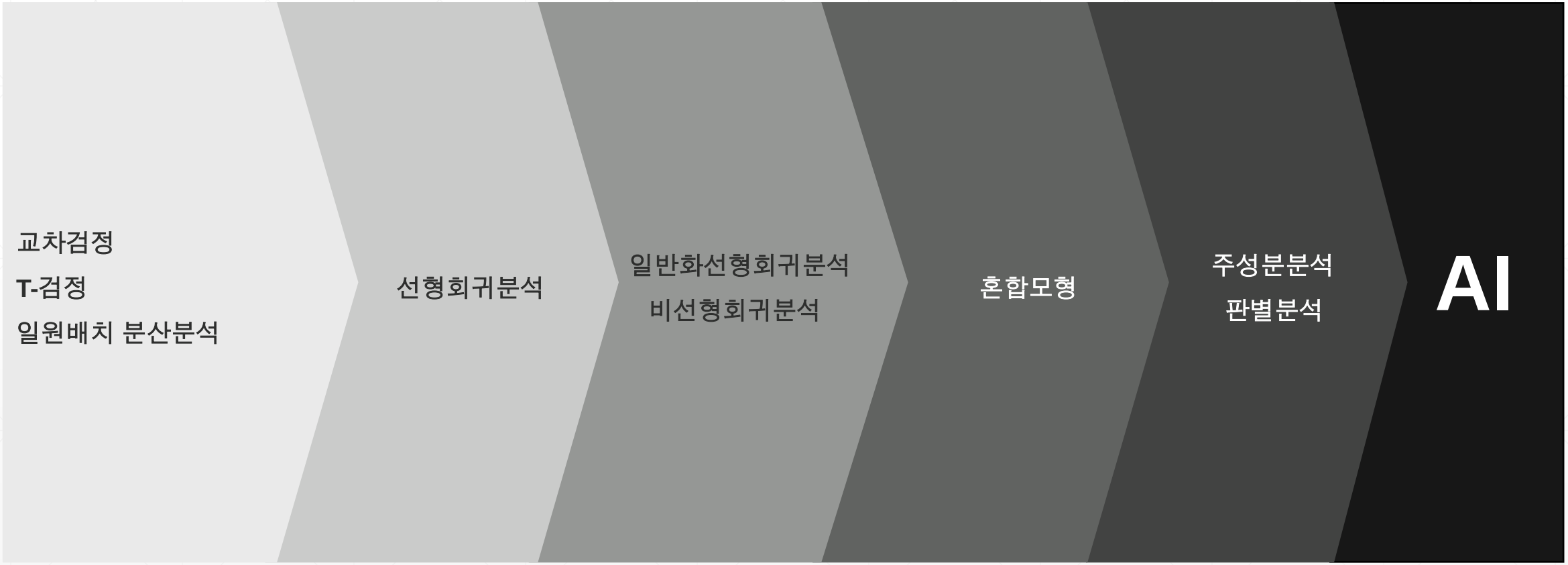

Dimension Reduction
Feature Extraction
• 선택한 변수 해석 어려움
• 변수간 상관관계 고려 가능
Feature Selection
• 선택한 변수 해석 쉬움
• 변수간 상관관계 고려 어려움
Supervised
Unsupervised
Supervised
Unsupervised
PLS(부분 최소제곱법)
PCA
AutoEncoder
Wavelets Transforms
Information Gain
Stepwise(Forward, Backward)
Lasso Regression
Genetic Algorithm
PCA Loading

2) 다변량분석 방법

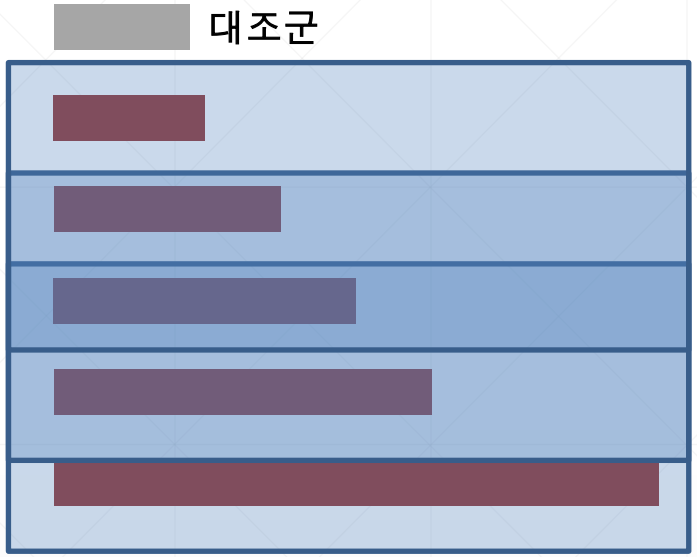


2) 다변량분석 방법

단일 분석 → 다변량 분석



4. 결론



BMI 22와 22.3 이 차이가 있다고?

P-value 에 대한 과한 맹신!!!!

머리에 이가 있으면 ……

…… 건강하다.

화재현장에 파견된 소방관들의 수가 많으면 …… 화재 손해액이 크다.

단일 요인으로만 해석

✓ 심프슨의 역설

- 하위 부분집단들에서 나타나는 결과와 이들을 결합한 전체집단에서 나타나는 결과가 상반되는 현상
- 1973년 버클리 대학원 입학허가에 관한 자료

	합격자	지원자	합격률
남자	1,400	2,691	52.0%
여자	722	1,835	42.1%

✓ 심프슨의 역설

- 하위 부분집단들에서 나타나는 결과와 이들을 결합한 전체집단에서 나타나는 결과가 상반되는 현상
- 1973년 버클리 대학원 입학허가에 관한 자료

[남자]

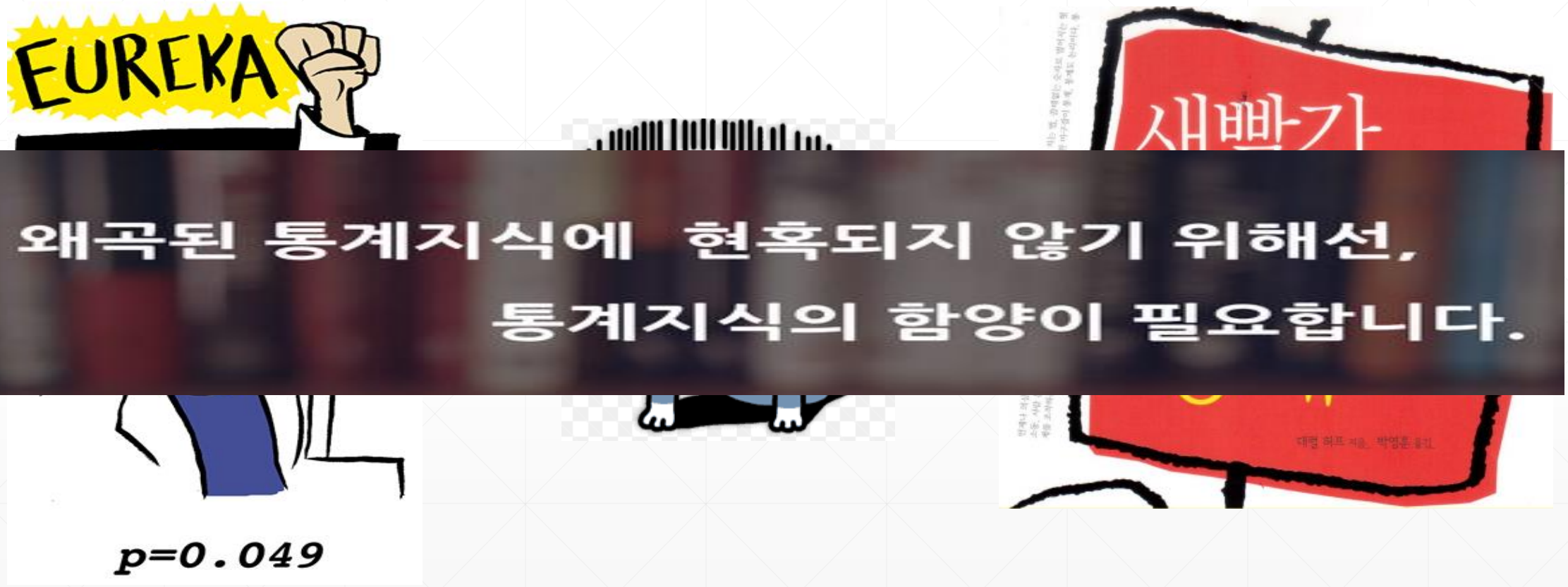
지원분야	합격자	지원자	합격률
A	512	825	62.1%
B	353	560	63.0%
C	120	325	36.9%
D	138	417	33.1%
E	53	191	27.8%
F	224	373	60.1%

[여자]

지원분야	합격자	지원자	합격률
A	89	108	82.4%
B	17	25	68.0%
C	202	593	34.1%
D	131	375	34.9%
E	94	393	23.9%
F	239	341	70.1%

4. 결론

4. 결론



Q & A ?

감사합니다.