통계학의 이해

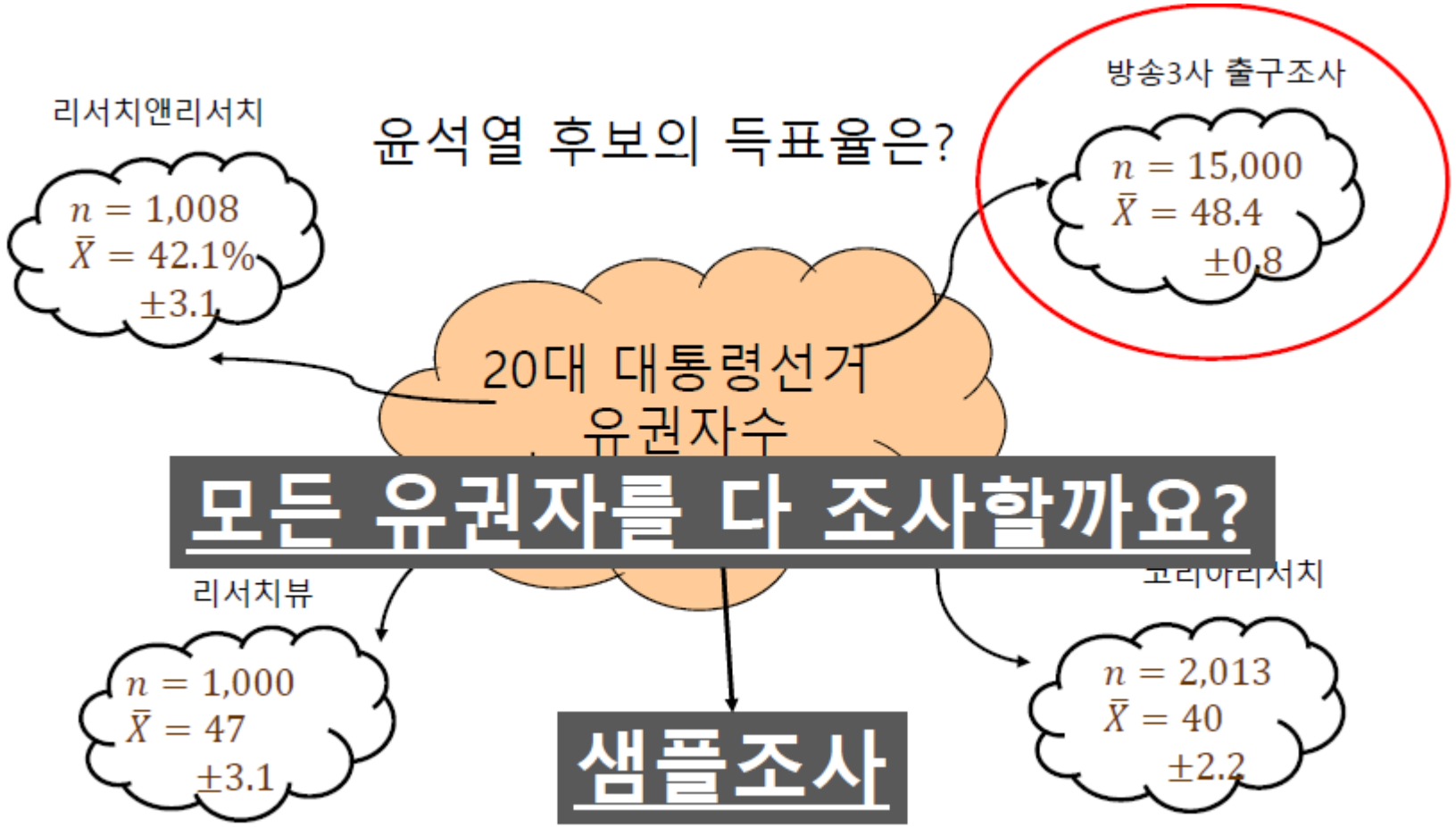
보건빅데이터통계분석

이새봄
삼육대학교 SW융합교육원

통계학이란

여론조사

- 20대 대통령 선거
  - 2022.3.1~



https://www.nesdc.go.kr/

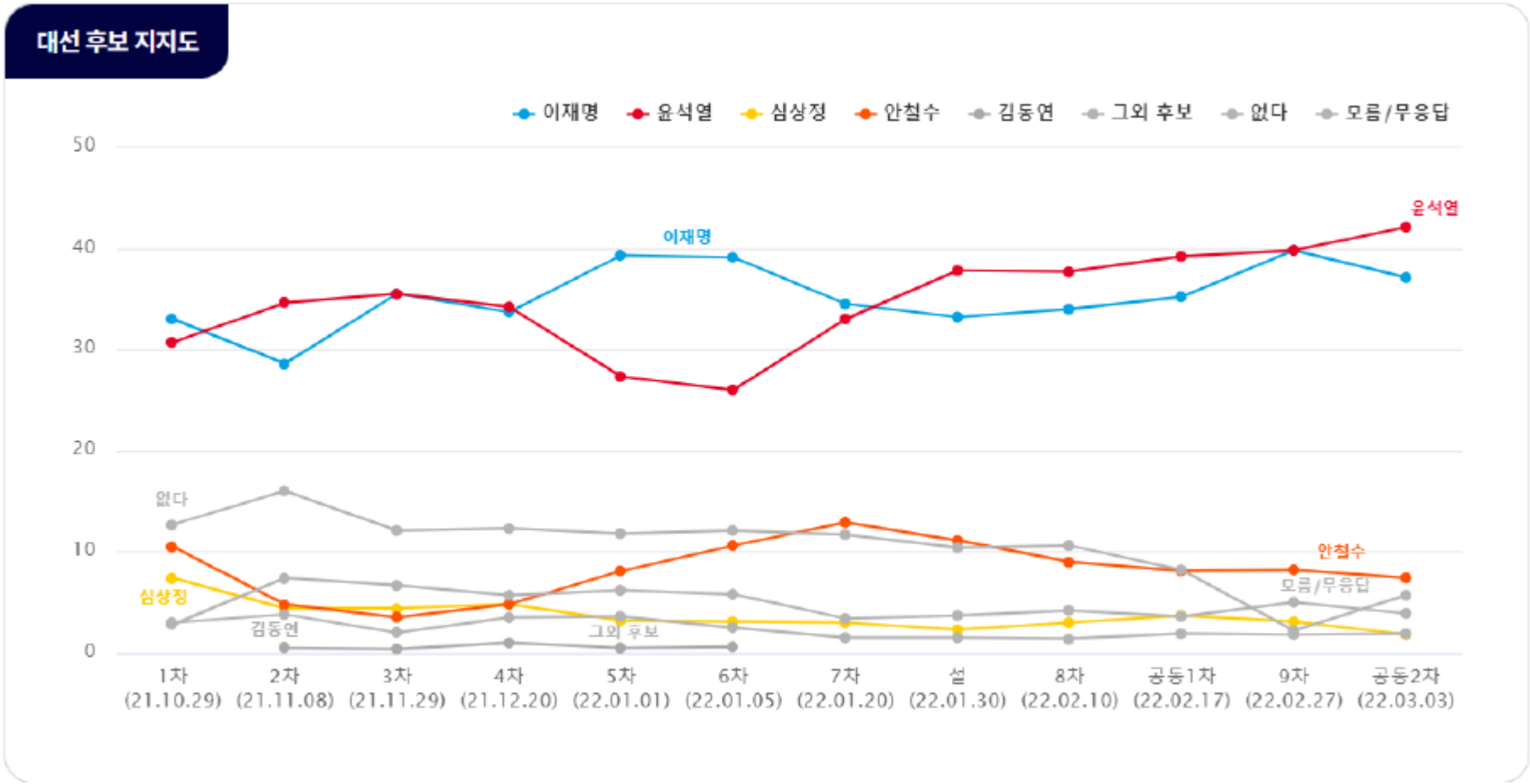
여론조사

지상파방송3사 공동 2차 여론조사 조사개요

• 조사의뢰기관	KBS, MBC, SBS
• 조사기관	입소스, 코리아리서치, 한국리서치
• 조사지역	전국
• 조사기간	2022년 03월 01일 ~ 03월 02일(2일간)
• 조사대상	전국에 거주하는 만18세 이상 성인남녀
• 조사방법	국내 통신 3사가 제공하는 휴대전화가상(안심)번호(100%)를 이용한 전화면접조사
• 표본크기	2,003명
• 피조사자 선정 방법	성/연령/지역별로 피조사자를 할당
• 응답률	24.9% (총 8,037명과 통화하여 그 중 2,003명 응답 완료)
• 가중치 부여방식	지역별, 성별, 연령별 가중치 부여(셀가중) (2022년 1월말 행정안전부 주민등록인구통계 기준)
• 표본오차	95% 신뢰수준에서 ±2.2% point

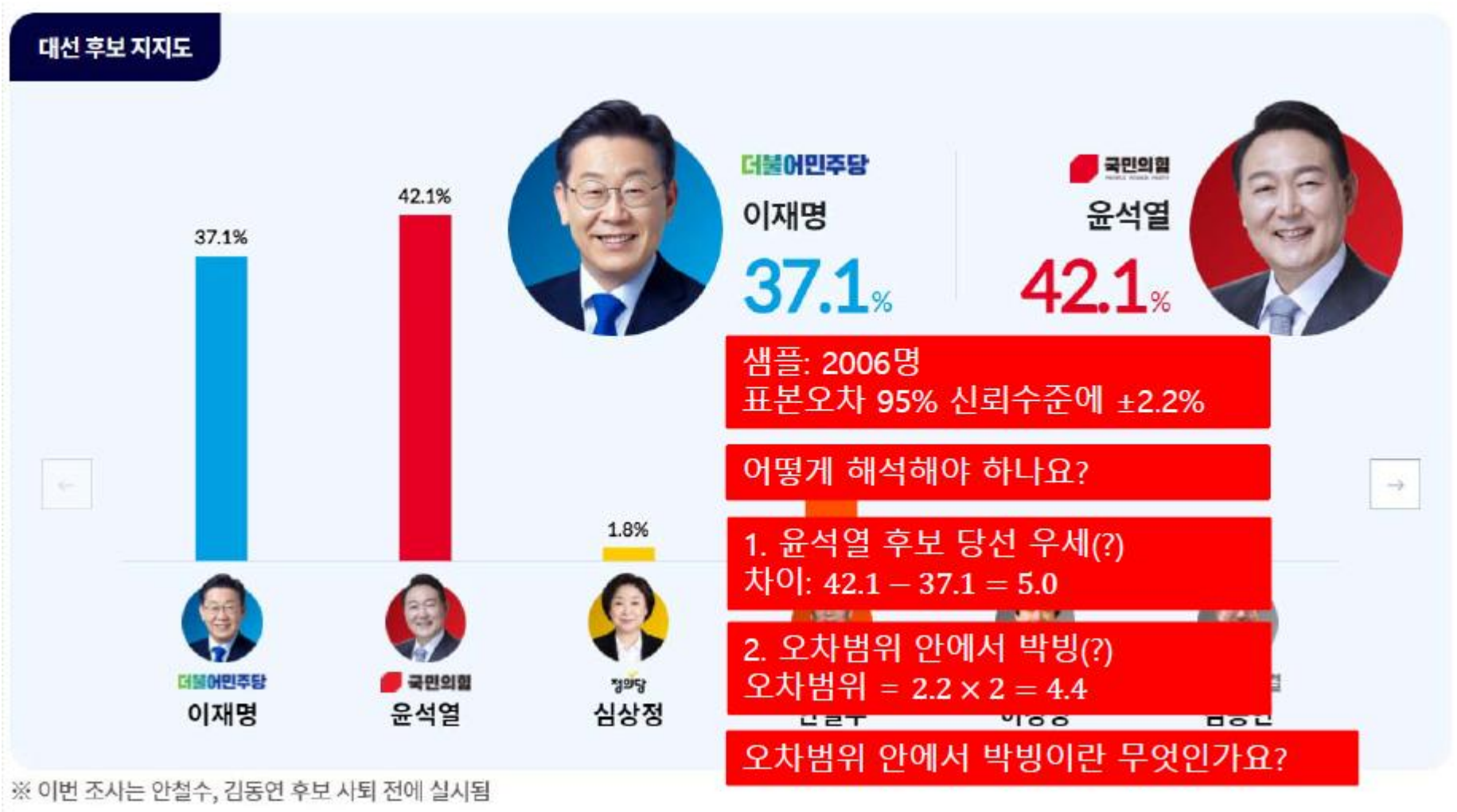
https://news.kbs.co.kr/special/election2022/president/poll_latest.html

여론조사



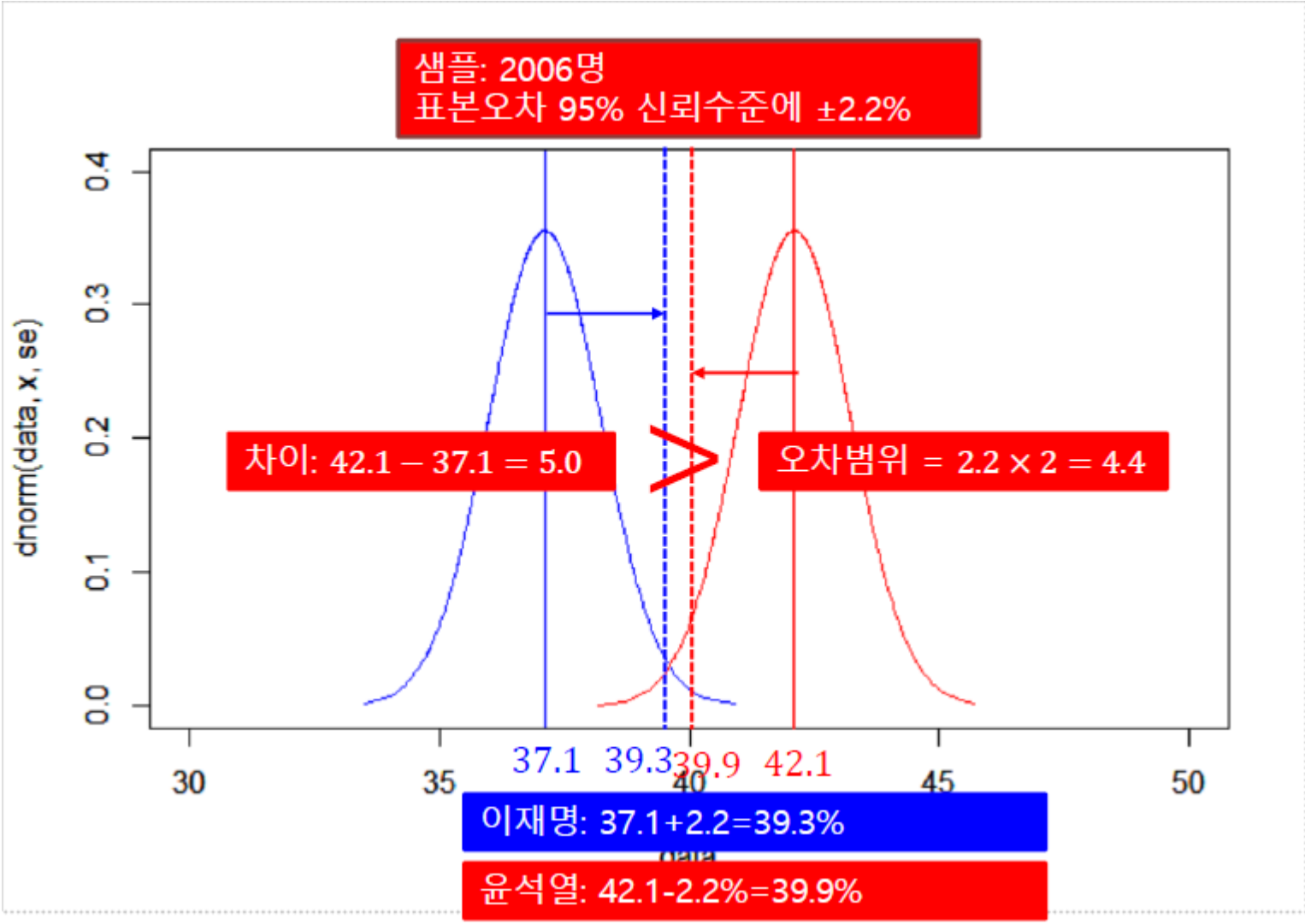
https://news.kbs.co.kr/special/election2022/president/poll_latest.html

여론조사



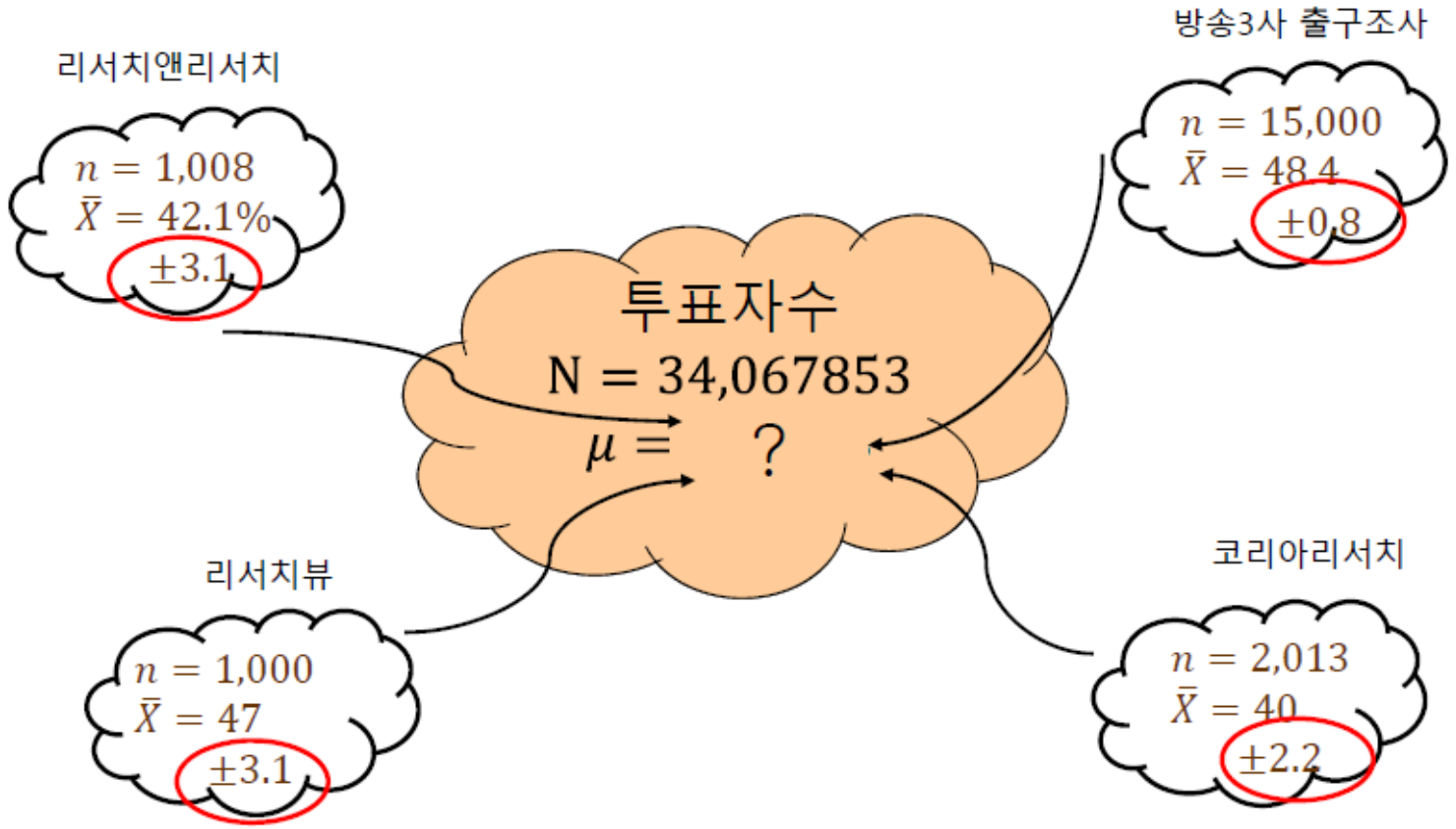
https://news.kbs.co.kr/special/election2022/president/poll_latest.html

여론조사


샘플: 2006명
표본오차 95% 신뢰수준에 ±2.2%
차이: 42.1 – 37.1 = 5.0
>
오차범위 = 2.2 × 2 = 4.4
dnorm(data, x, se)
0.0
0.1
0.2
0.3
0.4
30
35
40
45
50
37.1
39.3
39.9
42.1
이재명: 37.1+2.2=39.3%
윤석열: 42.1-2.2%=39.9%

여론조사

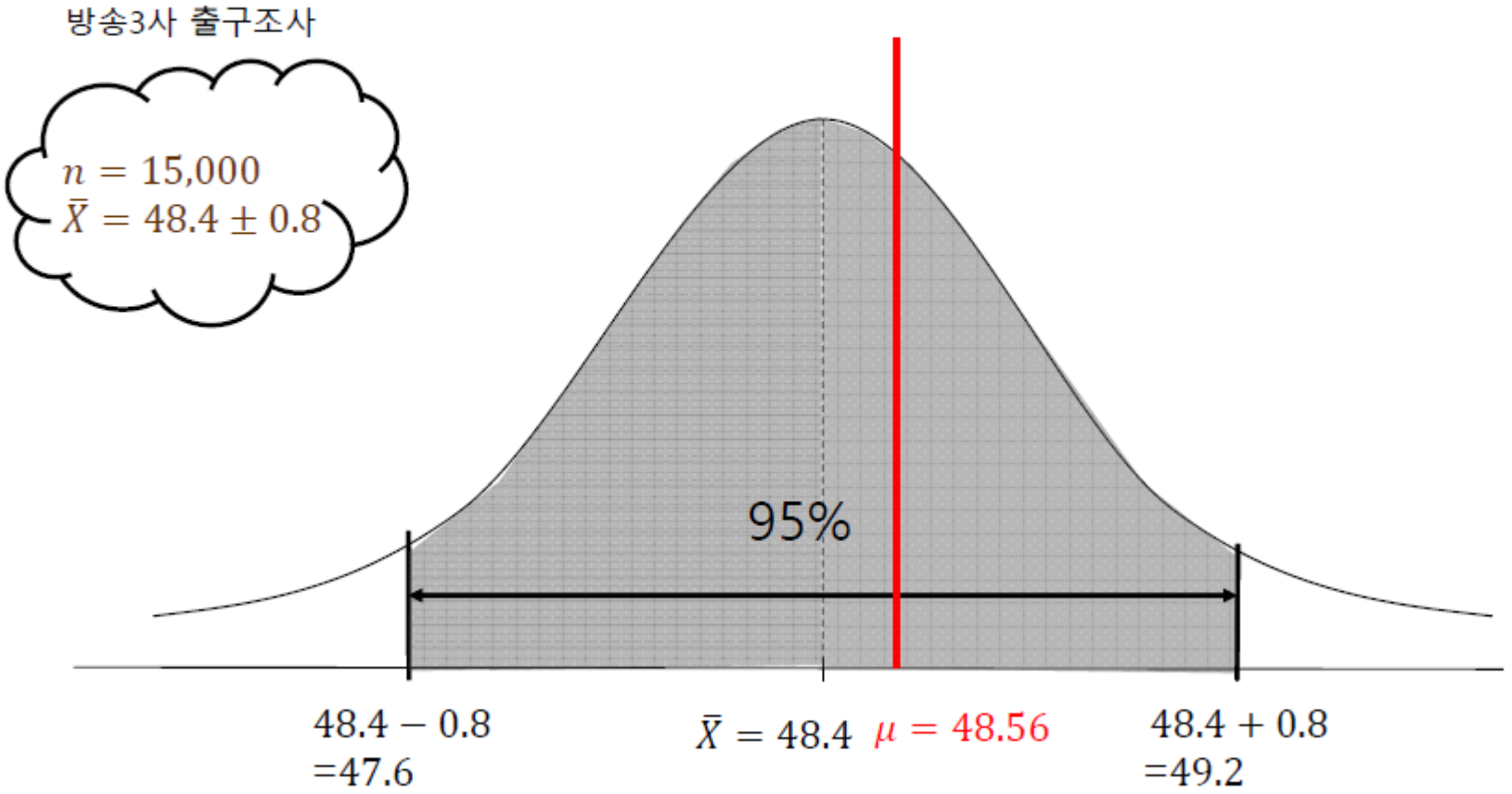
- 20대 대통령 선거(최종결과)



표본오차가 무엇인가요?

표본오차(오차범위)

- 신뢰구간: 모수가 있을 범위
  - 통계에서 많이 사용하는 신뢰수준: 95%



표본오차(오차범위)

- 조사일시: 2022.2.13 ~ 19

조사기관	이재명	윤석열	응답수	표본오차
서던포스트	31.4%	40.2%	1,001	±3.1%
리얼미터	38.7%	42.9%	3,043	±1.8%
칸타코리아	32.2%	41.3%	1,012	±3.1%
한국갤럽	34%	41%	1,007	±3.1%
리서치뷰	39%	48%	1,000	±3.1%
KSOI	43.7%	42.2%	1,002	±3.1%
한국리서치	36.9%	42.4%	1,000	±3.1%

응답수	100	500	1,000	2,000	5,000	10,000	20,000
오차범위	9.8	4.4	3.1	2.2	1.4	1.0	0.7

http://www.mediatoday.co.kr/news/articleView.html?idxno=302445

오차범위와 통계

- K대학 통계학 수업 수강생 100명 몸무게를 조사
  - 평균, 표준편차

$$\mu = 56.78$$
$$\sigma = 6.80$$

  - 몸무게가 55-60일 확률은?

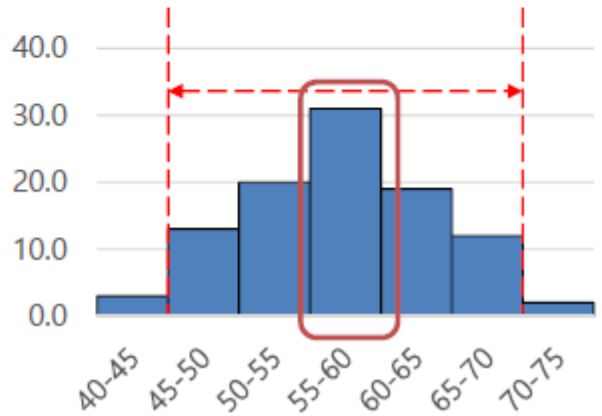
$$P(55 \le X \le 60) = 0.31$$

  - **몸무게 평균을 중심으로 95%확률로 예측할 수 있는 몸무게의 범위는?**

$$0.95 = P(45 \le X \le 70)$$

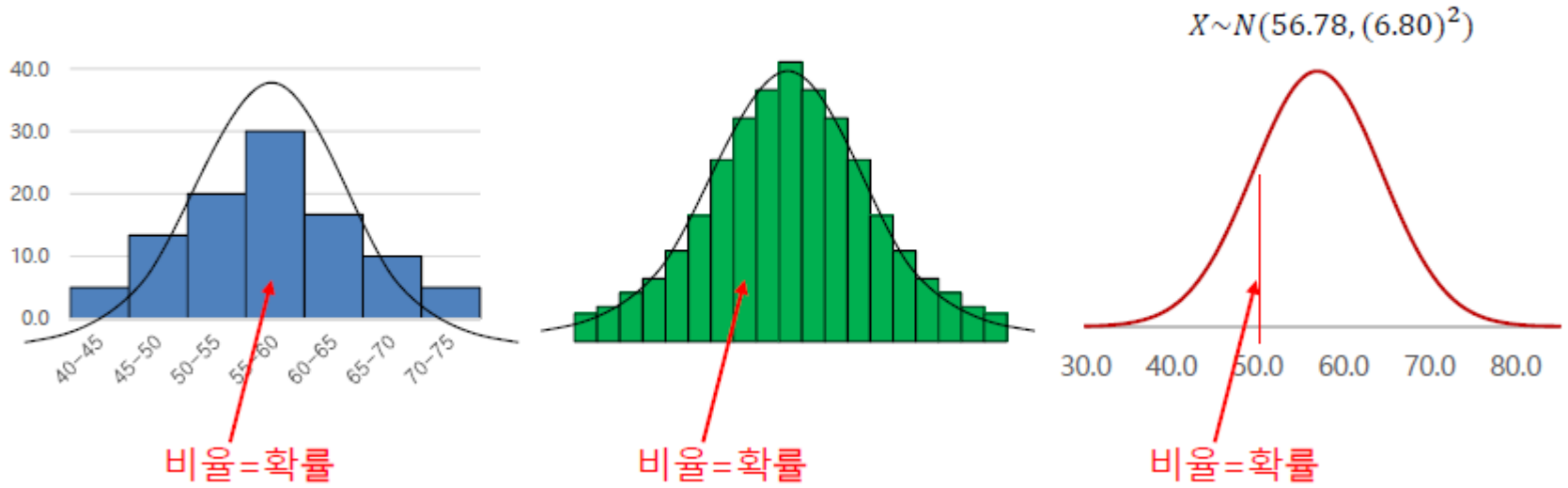
  - **고등학교 수학은 몸무게의 범위를 이용해 확률을 구하였으나, 이제는 확률을 이용해 범위를 구하는 것이 핵심**

X	빈도수	%	확률(p)
40-45	3	3	3
45-50	13	13	13
50-55	20	20	20
55-60	31	31	31
60-65	19	19	19
65-70	12	12	12
70-75	2	2	2
합계	100	1	1



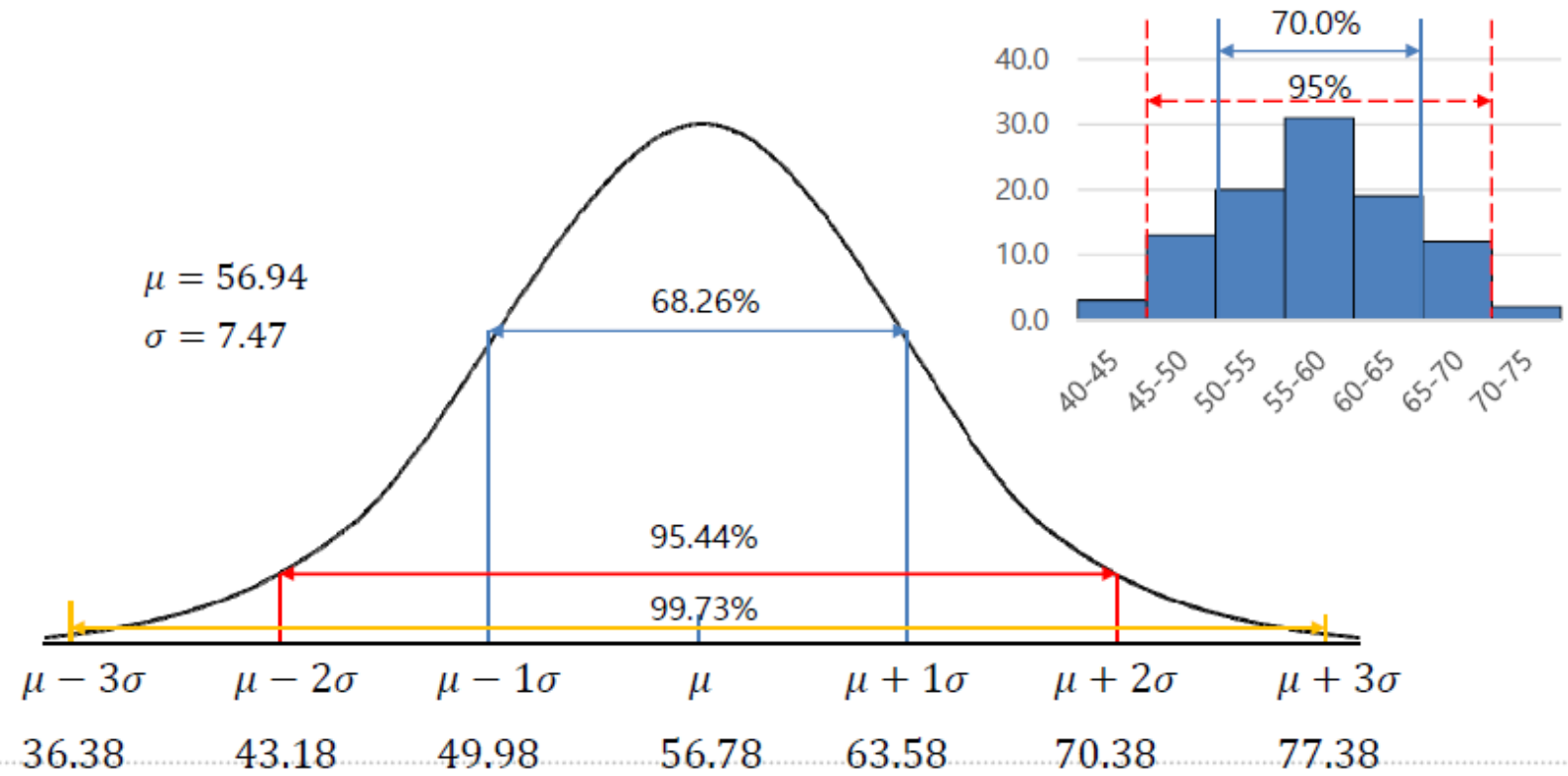
정규분포

- Normal Distriution



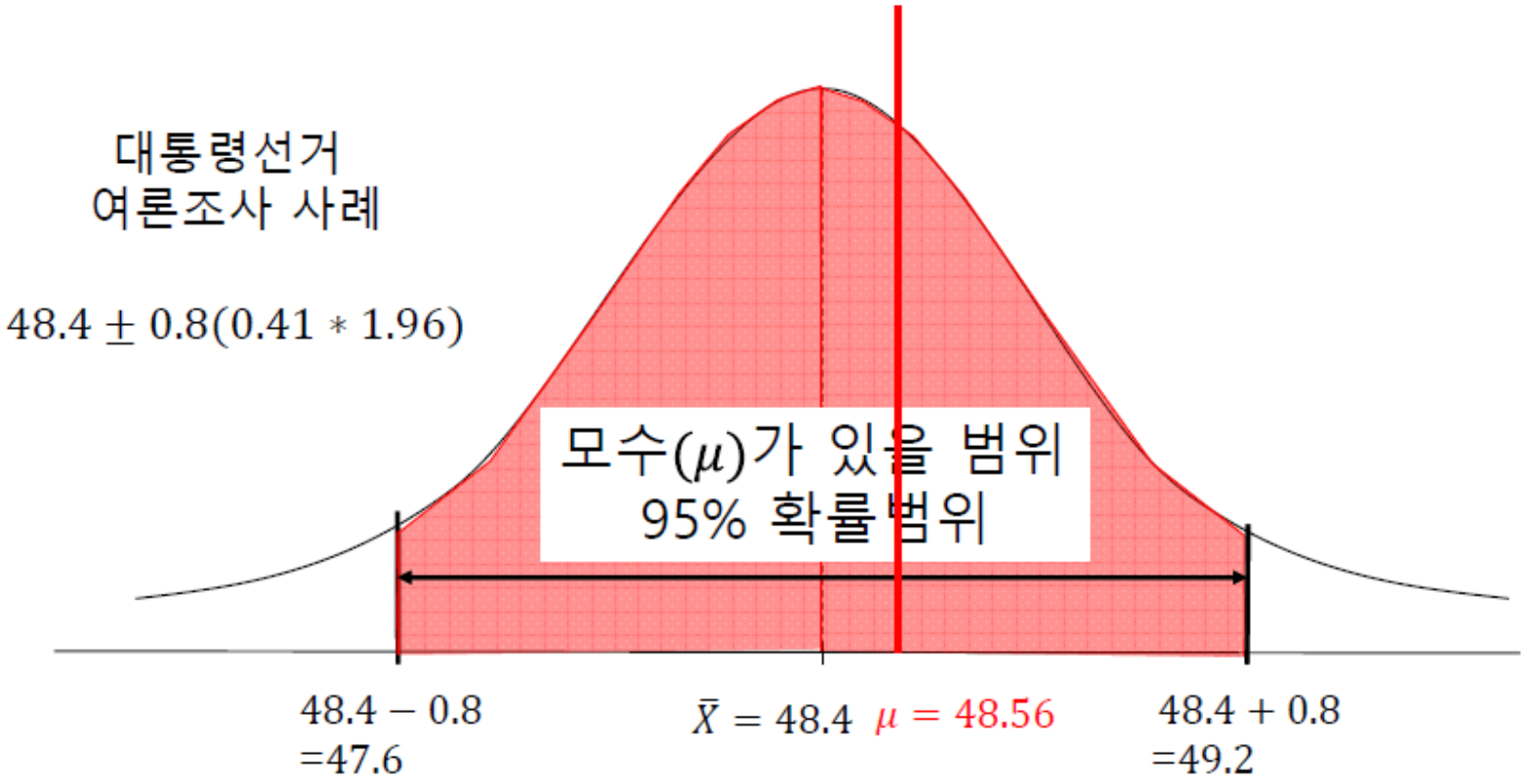
경험적 법칙

- Empirical Rule (경험적 법칙)
  - $k = 1,\ 68.26\%$ 이상의 데이터가 $\mu \pm 1\sigma$ 사이에 있음
  - $k = 2,\ 95.44\%$ 이상의 데이터가 $\mu \pm 2\sigma$ 사이에 있음
  - $k = 3,\ 99.73\%$ 이상의 데이터가 $\mu \pm 3\sigma$ 사이에 있음



확률과 오차 범위

- 통계적 방법론의 기초이론으로써 중요한 역할을 함
- 표본값을 이용해서 모수를 예측할 때 사용
- 모수가 있을 범위: 확률을 이용하여 모수를 추측



통계학이란

- 관심 대상인 모집단의 특성을 파악하기 위해
  - 문제: 모집단을 조사하기 어려움

- 모집단으로부터 관련된 일부 자료(표본)을 수집하고
  => 표본추출 (Sampling)

- 수집된 표본의 자료를 요약하여 표본의 특성을 파악하고
  => 기술통계학(descriptive statistics)

- 표본의 자료를 이용하여 모집단의 특성에 대한 확률을 이용해 추론하는 학문
  => 추론(추측)통계학(inferential statistics)

통계학이란

- 통계에 근거하여 의사결정을 위한 분석 기술과 절차를 다루는 학문
- 데이터를 수집하고, 과학적으로 분석하여 의사결정에 활용하는 학문
- ‘집단 현상을 수량적으로 관찰하고 분석하는 방법 ' 을 연구하는 학문
- 특정 문제가 주어졌을 때, 합리적인 의사결정을 내리기 위해 통계기법을 도구로 사용하는 학문

기술통계학

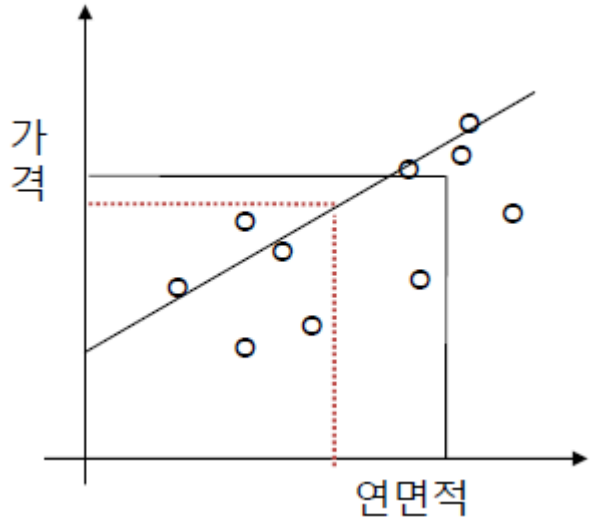
- 기술통계학(descriptive statistics)
  - 조사 및 측정된 자료를 통해 그 자료가 가지고 있는 특징을 수치, 표, 그래프로 정리하는 과정
- 범주형 자료
  - S대학 경영학부의 남자와 여자의 성비

변수	항목	빈도	%
성별	남자	101	57.7
	여자	74	42.3
Total		175	100.0

- 수치형 자료
  - S대학 2학년 학생의 키: 평균 175 ±5cm

추론통계학

- 관심대상 전체로부터 일부의 샘플을 추출, 분석하여 그 결과로부터 전체 모집단에 대한 특성을 예측
  - 사례1) 아이스크림의 용량은 320g을 판매해야 된다. A매장에서는 320g을 팔고 있는지 표본 100개를 가지고 확인해 보자.
  - 사례2) 주택의 연면적에 따른 주택가격을 예측해보자.



모집단과 전수조사

- 모집단(population)
  - 관심 있는 연구대상 전체의 집합
  - 무한모집단: 모집단의 크기가 무한한 경우 (전세계 인구, 자판기 커피)
  - 유한모집단: 모집단의 크기가 유한한 경우 (삼육대학교 재학생)

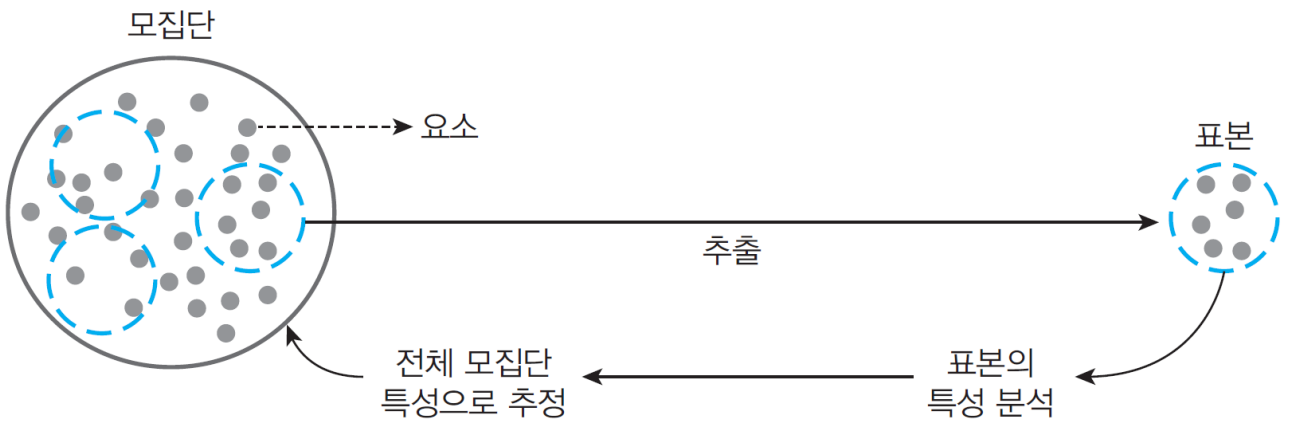
- 전수조사
  - 관심 있는 모집단 전체를 조사하는 경우로서 주로 모집단의 규모가 작을 경우에 실시

- 전수조사의 어려움
  - 조사불가능: 모집단 전체를 대상으로 조사하기는 불가능
  - 시간과 비용: 모집단을 다 조사하는 데는 많은 시간과 비용 소요

- 해결책: 전수조사 → 표본조사

표본과 표본조사

- 표본 (Sample)
  - 추출된 모집단 구성의 일부
  - 모집단의 일부분으로부터 얻어진 관측 값의 집합



- 표본조사
  - 모집단에서 추출된 일부부인 표본을 가지고 하는 조사
  - 수집방법: 실험, 조사, 출판자료

모수와 통계량

- 모수(parameter):
  - 모집단에 대한 수치 특성값
  - 모집단의 특성을 나타내는 양적인 측도로서 주어진 모집단을 따르는 고유의 상수 (상수=모집단은 진실된 하나의 값임)
  - 모평균(μ), 모표준편차(σ)
  - 예) 우리나라 고등학교 사교육비 평균

- 통계량(statistic):
  - 표본에서 얻은 수치 특성값
  - 표본의 특성을 나타내는 양적인 측도로서 모집단의 분포를 따르는 확률변수
  - (확률변수=표본에 따라 값이 변함)
  - 표본평균(X), 표본표준편차(s)
  - 예) 우리나라 고등학교 1학년 중에서 1000명만 뽑아 조사하여 얻은 평균 사교육비

표본오차와 통계적 추론

- **표본오차**(sampling error)
  - 모집단에서 표본을 추출해서 조사하기 때문에 모수와 표본 통계량 사이에 생기는 오차
  - 표본의 크기를 크게 함으로써 표본오차를 감소 → 통계학에서 표본의 크기를 크게 하라는 이유
  - 표본오차는 아무리 표본을 크게 해도 전수조사를 하지 않는 이상 존재
  - 표본오차의 허용범위를 확률로 구하는 것이 통계의 목적

- **통계적 추론**
  - 우리가 실제로 알고 싶은 것은 표본의 값(통계량: statistic)이 아니고 모집단의 값(모수: Parameter)
  - 통계학의 목적: 추론(Inference) → 표본에서 구한 값을 이용해 우리가 구하고자 하는 모집단의 값도 이럴 것이다라고 추론

표본조사 방법

- 확률추출(probability sampling)
  - 모집단에 속하는 모든 추출단위에 대해 사전에 일정한 추출확률이 주어지는
  - 표본추출법
  - 표본추출틀 존재
  - 단순확률추출(simple random sampling)
  - 계통추출법(systematic sampling)
  - 층화확률추출(stratified random sampling)
  - 집락추출법(cluster sampling)
- 비확률추출
  - 추출단위가 표본에 추출될 확률을 객관적으로 나타낼 수 없는 표본추출법
  - 편의표출(convenience sampling)
  - 할당추출(Quota sampling)
  - 포커스 그룹(Focus Groups)

표본조사 방법

- 기본용어
  - 기본단위(elementary unity) : 조사의 대상이 되는 가장 최소의 요소
  - 예)여론조사 : 개인, 가계조사 : 가구, 농작물조사: 일정 면적의 경지

- 추출틀(sampling frame) : 모집단에 속하는 모든 추출단위의 목록
  - 예)개인, 가구, 사업체 등의 명부, 문서철, 지도 등

- 목표모집단(target population): 관심을 갖고 특성을 알아보고자 하는 집단에 속하는 모든 기본단위들의 집합
  - 예) 삼육대학교 재학생 학부모

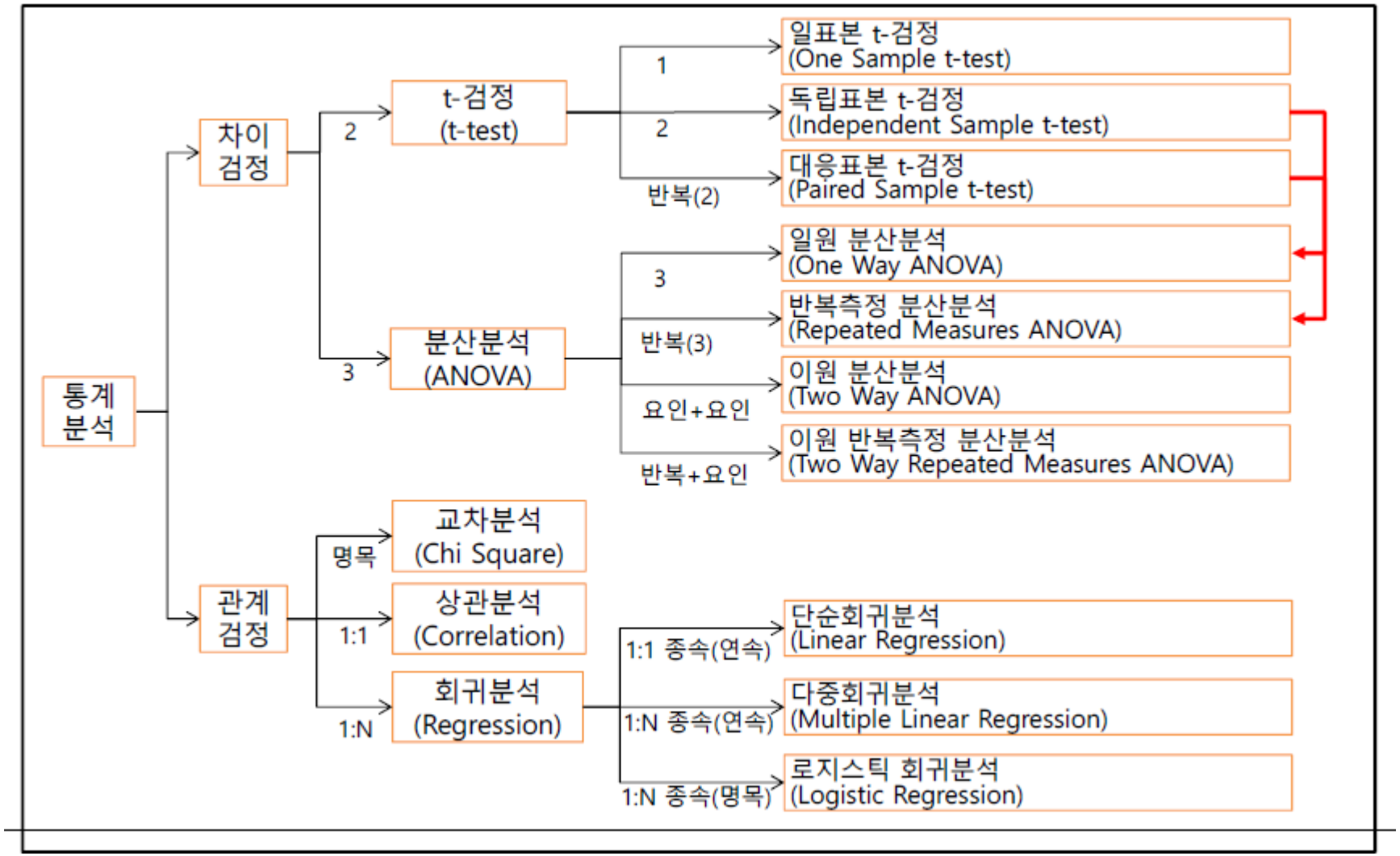
- 조사모집단(target population): 표본 추출틀을 통해 추출될 수 있는 기본단위들의 집합(실제 조사 가능한 집단)
  - 예) 전화조사 : 삼육대학교 재학생 학부모 중 전화번호가 있는 사람

통계 분석 방법

통계분석방법

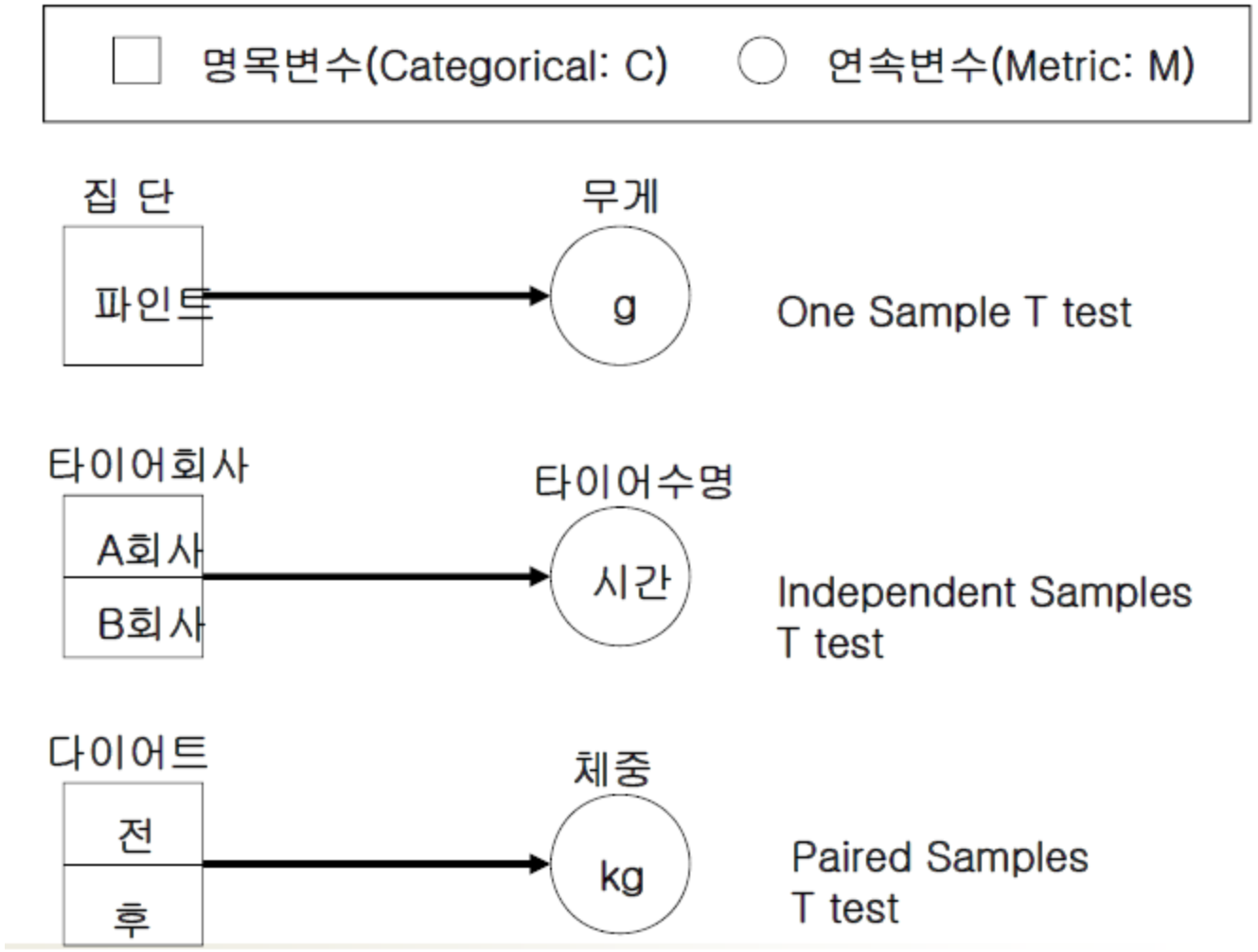
- 평균차이검정
  - 집단간 평균차이를 검정하는 방법
  - 평균검정(T-test), 분산분석(ANOVA), ANCOVA, MANOVA
- 관계검정
  - 변수와 변수의 관계를 검정
  - 상관분석, 회귀분석, 교차분석, 정분상관분석, 판별분석, 로지스틱회귀분석
- 신뢰도와 타당도
  - 신뢰도분석, 요인분석
- 기타
  - 군집분석, 다차원척도법, 생존분석, 데이터마이닝 기법등

통계분석방법



출처: 간호보건 통계분석, 한상숙, 이상철

차이검정


명목변수(Categorical: C)
연속변수(Metric: M)
집 단
무게
파인트
g
One Sample T test
타이어회사
타이어수명
A회사
B회사
시간
Independent Samples
T test
다이어트
체중
전
후
kg
Paired Samples
T test

차이검정

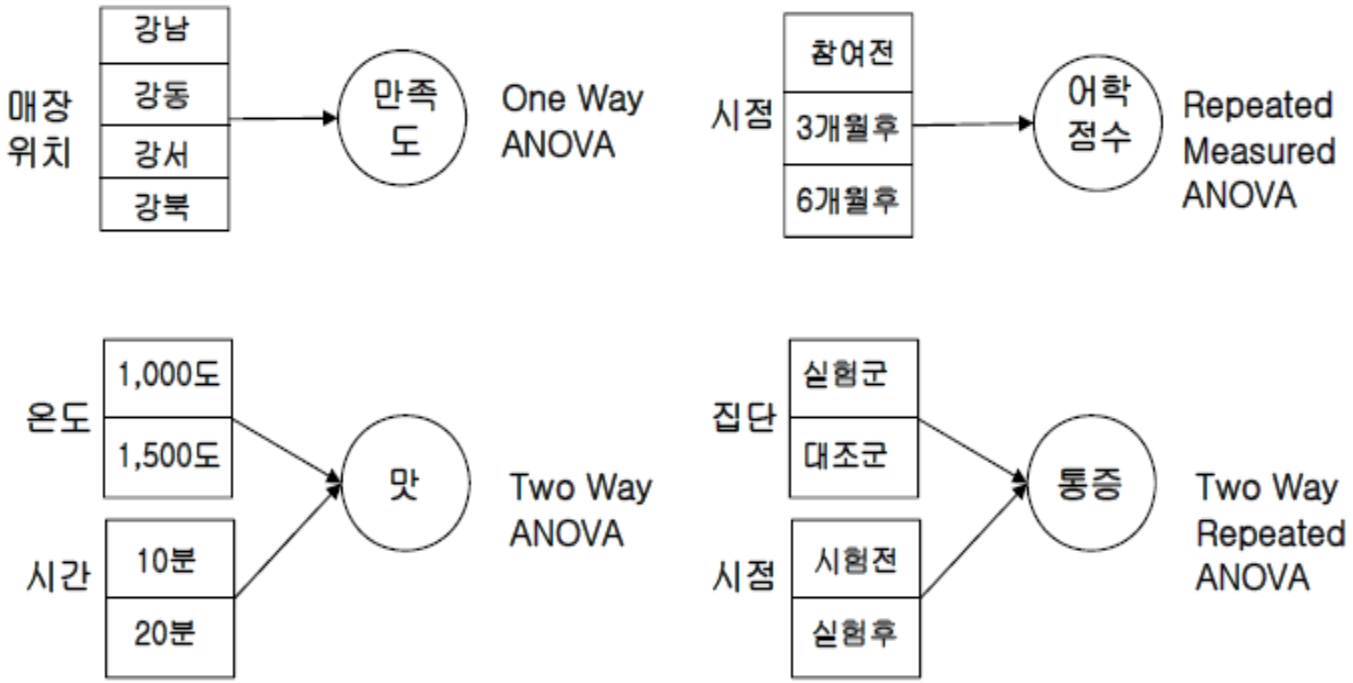
- 분산분석(ANOVA)
  - 집단이 3개 이상일 때
  - 일원배치 분산분석(One-Way ANOVA): 한 개의 집단구분변수
  - 이원배치 분산분석(Two-Way ANOVA) : 두 개의 집단 구분 변수를 동시에
  - 반복측정 분산분석(Repeated Measures ANOVA): 집단이 세 개이면서 반복 측정

- 평균분석(T-Test)과 분산분석(ANOVA)의 비교
  - 왜? 평균분석을 3번 안 할까?
  - 예) 1,2,3 이라는 집단
  - 1:2, 2:3, 1:3 → 3번 평균비교 VS 분산분석(ANOVA) 1번
  - 여러 번 평균검정(T-test)를 해주게 되면 1종 오류를 범할 확률이 증가

$$1-(1-\alpha)^t = 1-(1-0.5)^3 = 0.14$$

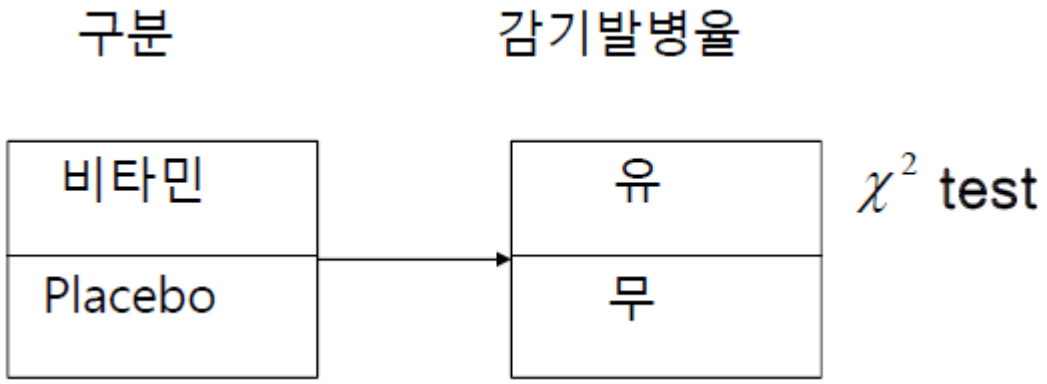
  - 3개 이상일 때는 모든 집단의 평균이 같은지를(1=2=3) 분산을 이용하여 분석
  - 분산분석을 실시한 후에 3 개의 집단 간에 평균이 틀리다고 밝혀지면 그때 비로서 각 집단간 3쌍의 평균을 비교
  - 이때도 Duncan, Tukey 등 1종 오류를 보정해 주는 방법을 이용해서 분석

차이검정


매장 위치
강남
강동
강서
강북
만족도
One Way ANOVA
시점
참여전
3개월후
6개월후
어학 점수
Repeated Measured ANOVA
온도
1,000도
1,500도
시간
10분
20분
맛
Two Way ANOVA
집단
실험군
대조군
시점
시험전
실험후
통증
Two Way Repeated ANOVA

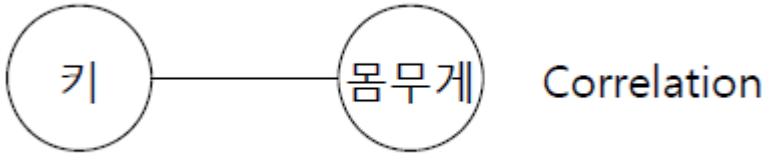
관계검정

- 관계검정
  - 변수와 변수의 관계를 검정
  - 상관분석(Correlation Test): 연속변수 + 연속변수
  - 회귀분석(Regression): 연속변수 + 연속변수
  - 교차분석(test): 질적변수 + 질적변수
- 교차분석
  - 2개의 질적변수



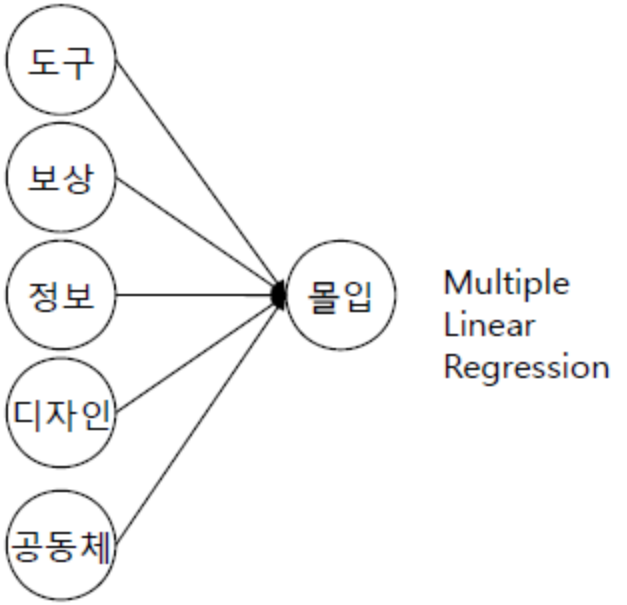
관계검정

- 상관분석(Correlation Test)
  - 상관관계: 두 변수가 서로 동등한 입장에서 관계를 분석
  - 예) 몸무게가 많이 나가면 허리둘레도 크고, 반대로 허리둘레가 크면 몸무게도 많이 나감
  - 편상관분석(Partial Correlation) : 중간에 다른 변수의 영향력이 있을 때 이를 통제

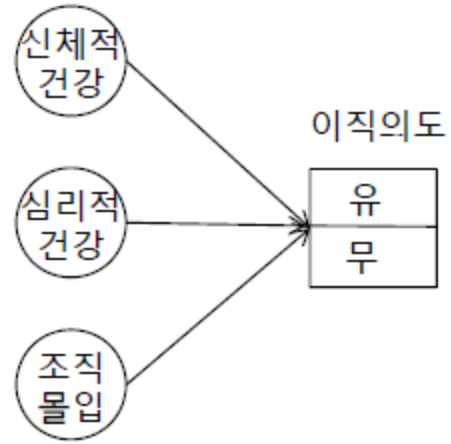


- 회귀분석(Regression)
  - 인과관계: 하나의 변수가 원인이 되어 다른 변수(들)에 영향을 미치는 관계

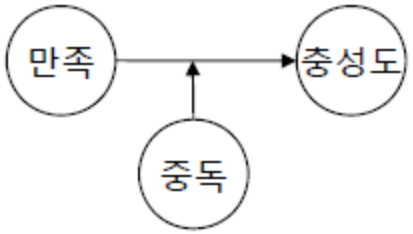
관계검정



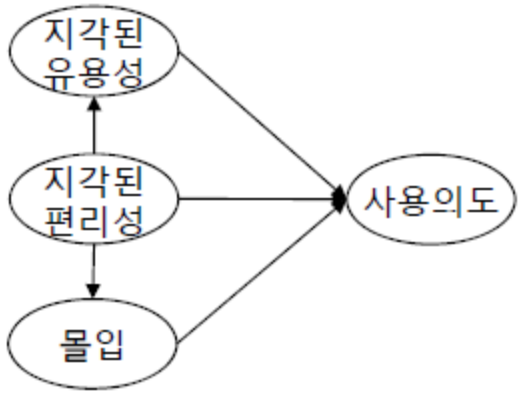
Multiple
Linear
Regression



Logistic
Regression



Hierarchical
Regression



Structural
Equation
Model

차이검정

- 평균차이검정
  - 집단간 평균차이를 검정하는 방법
  - 질적변수 1개 (집단구분) + 연속 변수 1개 (평균)
  - 평균검정(T-test): 집단이 2개 미만 일 때
  - 분산분석(ANOVA): 집단이 3개 이상일 때
- 평균검정(t-검정)
  - 단일 표본 t-검정(One Sample T test): 하나의 집단의 평균이 얼마인지를 검정
  - 독립 두 표본 t-검정(Independent Samples T test): 흡연집단과 비흡연집단으로 나누어서 이 두 집단간의 제태기간의 차이가 있는지를비교
  - 대응표본 t-검정검정(Paired Samples T test): 하나의 집단이지만 이 집단을 처리전과 처리 후로 두 개로 나누어서 비교(다이어트효과)

통계적 추론의 종류

- 추정(Estimation)
  - 표본의 평균과 표준오차(SE)를 구해서 모수의 범위를 구하는 것
  - 신뢰구간: 일정한 확률범위 내에서 모수의 값이 포함될 가능성이 있는 범위
  - 90%, 95%, 99% 의 확률 값 중에서 95%
  - 종류: 점추정, 구간추정
  - 사례: $\mu = 295.4 \pm 7.26, [288.14, 302.66]$

- 가설검정
  - 모수는 얼마이다라고 정하고 그것이 맞는지 틀리는지를 검증하는 방법
  - 유의수준 (α) : 모수와 통계량의 차이가 커서 확률적으로 가설을 기각할 수 있는 값
  - 1%, 5%, 10%의 값 중에서 5%
  - 종류: 귀무가설 (Null Hypothesis), 연구가설 (Alternative Hypothesis)
  - 사례: $t_{cal} = -12.25 < t_{critical} = -1.984,\ p - value = 0.000 < \alpha = 0.05$

가설검정

- 가설 검정(Hypothesis Test)
  - 모집단 모수의 값을 설정하고(가설 설정), 표본 통계치를 통해 확률적으로 진위를 판정하는 과정

- 가설유형
  - 추론통계에서는 귀무가설과 대립가설을 세우고, 귀무가설이 기각됨을 통해 본래 알아보고자 한 대립가설이 통계적으로 유의한 것인지를 확인
  - H0: 귀무가설(Null Hypothesis, 영가설)
    - 기존에 알려져 있는 사실(status quo), 통계적 검정 대상
    - 알아보고자 하는 내용과 반대되는 내용으로서 '두 변수 사이에 관계가 없다' 또는 '두집단은 차이가 없다'와 같이 부정적인 형태로 진술
  - H1: 대립가설(연구가설)
    - 새로운 사실, 현재 믿음에 변화가 있는 사실, 뚜렷한 증거로 입증하려고 하는 주장
    - 알아보고자하는 내용으로서 '검정할 가설의 내용에는 차이가 있다' 또는 '효과가 있다' 와 같이 진술

- 사례) H자동차에서 만든 새로운 하이브리드 차량의 연비는 $16.5km/l$로 알려져 있다. 과연 진짜로 $16.5km/l$ 인가를 검증

통계적 가설검정

- 통계적 가설검정: 귀무가설을 받아들일 것인지, 아니면 기각(reject)할 것인지를 검증
- 통계분석 귀무가설 설정(무죄주의 원칙)

구분	방법	귀무가설
차이검정	$t-test$, ANOVA	$H_0 : \mu_1 = \mu_2$
관계검정	regression	$H_0 : \beta = 0$

- 가설검정은 $H0$가 진실인지를 검증함

가설 검정의 종류

귀무가설(H_0)

$H_0: \mu = 16.5$

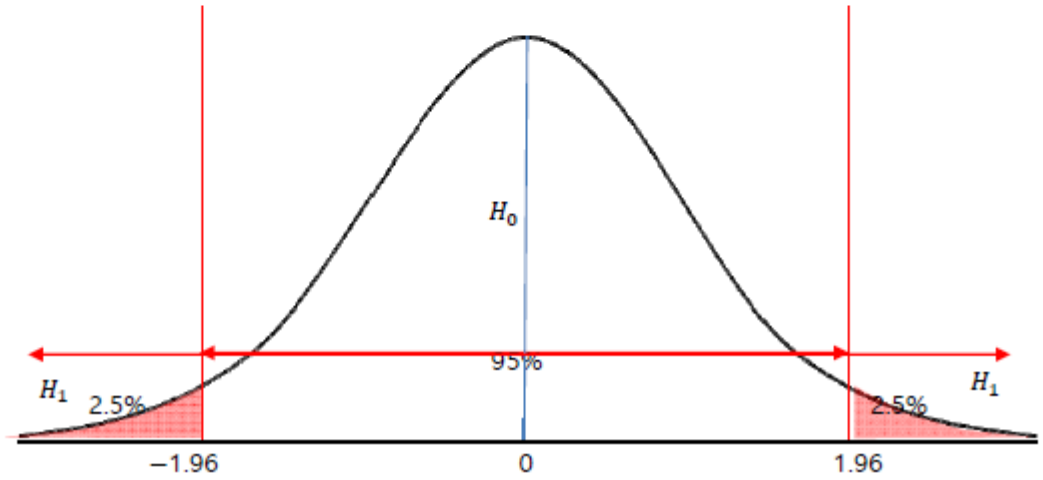
대립가설(H_1)

$H_1: \mu \neq 16.5$ 양측검정(two-sided test)

$H_1: \mu > 16.5$ 우측검정(right-sided test)

$H_1: \mu < 16.5$ 좌측검정(left-sided test)

양측검정(two-sided test)



가설 검정의 종류

귀무가설(H_0)

$H_0: \mu = 15$

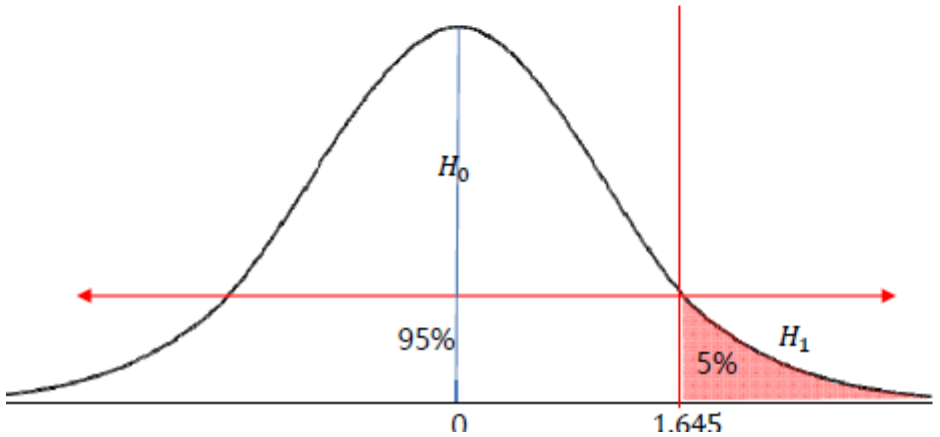
대립가설(H_1)

$H_1: \mu \neq 15$ 양측검정(two-sided test)

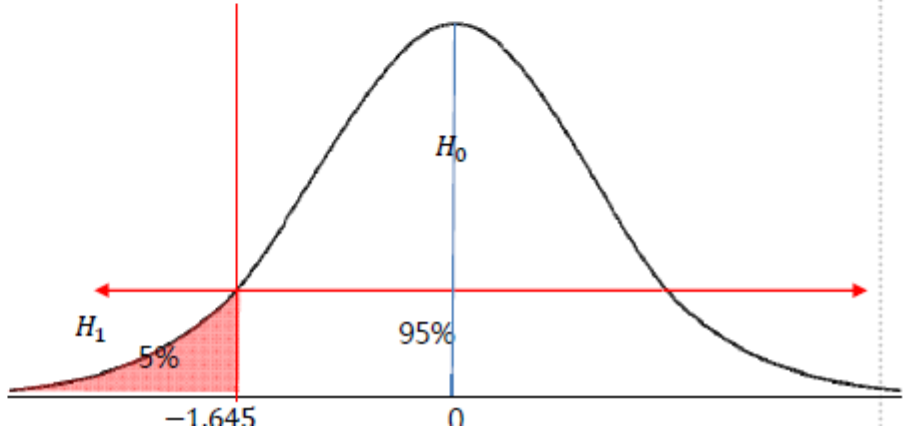
$H_1: \mu > 500$ 우측검정(right-sided test)

$H_1: \mu < 5$ 좌측검정(left-sided test)

H_1:우측검정(right-sided test)



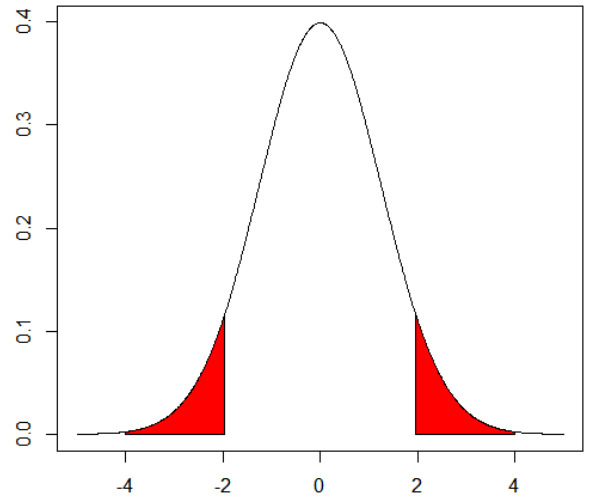
H_1:좌측검정(left-sided test)



귀무가설의 기각 수준

- 가설을 검증할 때는 귀무가설이 채택될 가능성을 본다. 보통 신뢰도 95% 수준으로 본다.
- 따라서 귀무가설이 채택될 가능성이 평균을 중심으로 95% 구간에 있으면 귀무가설은 채택되고, 대립가설은 기각된다. 즉, 본래 알아보고자 한 내용은 통계적으로 유의하지 않게 된다.

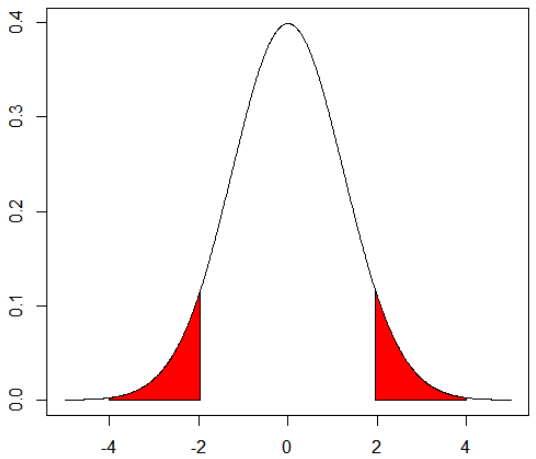
- 반대로 귀무가설이 채택될 가능성이 왼쪽 0.25%, 오른쪽 0.25% 구간에 속하게 되면 귀무가설은 기각되고,
- 대립가설이 채택된다. 즉, 본래 알아보고자 한 내용이 통계적으로 유의하다는 의미이다.

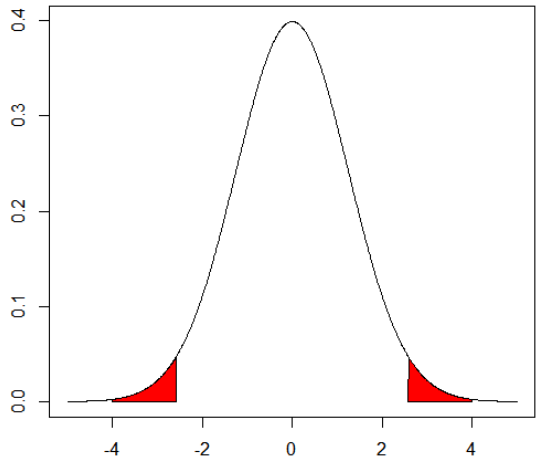


귀무가설의 기각 수준

p-value

- 이 때 사용된 기준을 유의수준이라고 하고 이 경우에는 5% 즉, 0.05가 귀무가설이 기각되고 대립가설이 채택될 수 있는 기준이 된다.
- 분석결과 가설이 속할 확률인 유의확률 **p-value**가 **0.05**보다 크면 귀무가설을 채택하고, **0.05**보다 작으면 대립가설을 채택한다.
- 보다 유의한 수준을 보기 위해서는 p-value가 0.01보다 작은지를 본다





Q&A