

# Lecture Note 04



## Fall, 2021

# Syllabus

Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	
4	9/27(월)	<b>Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증</b>	과제#2
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#3
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 콘텐츠 수집)	과제#4
7	10/18(월)	Predictive Analytics I – Linear regression & Logistic Regression	
8	10/25(월)	Predictive Analytics II – Clustering & Latent Class Analysis	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Tree-based Model and Bagging (Random Forest)	
10	11/8(월)	Predictive Analytics IV – Association Rules	
11	11/15(월)	Supplementary Topic II - 외부 데이터 수집 (동적 콘텐츠 수집)	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	12/20(월)	<b>Final Presentation</b>	

## Lecture 4-1

데이터 요약을 통해  
특성 파악하기

# 다음의 데이터를 살펴보자.

Chicago의 일 평균 초미세먼지(pm2.5) (1987년 1월 1일 ~ 2005 12월 31일)



최소값 (Min)	1.70
1분위수 (1 <sup>st</sup> Quantile)	9.70
중앙값 (Median)	14.66
3분위수 (3 <sup>rd</sup> Quantile)	20.60
최대값 (Max)	61.50
평균 (Mean)	16.23
표준편차 (Standard Deviation)	8.7

위 숫자만 보고 파악할 수 있는 시카고 시의 초미세먼지 특성은 무엇일까 ?

# 매출액의 평균을 비교해보자

어느 백화점 매출액이 평균적으로 가장 높을까?



2300만원



2350만원



2400만원

# 매출액의 평균을 비교해보자

어느 백화점 매출액이 평균적으로 가장 높을까?



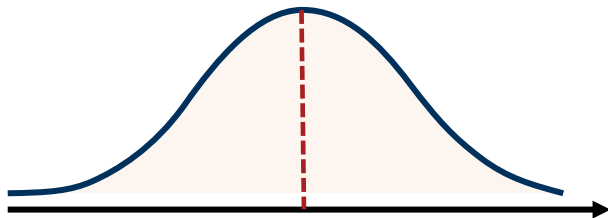
2300만원



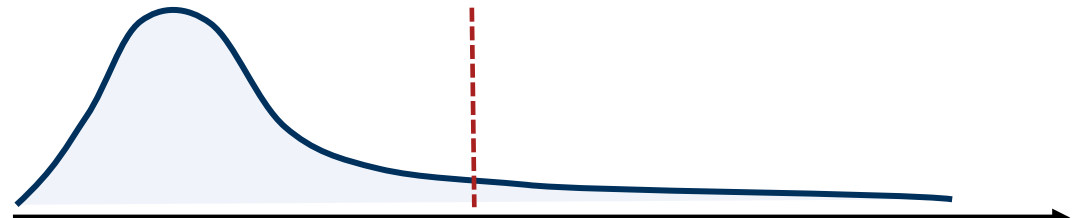
2350만원



2400만원



2300만원



2350만원 or 2400만원

# 위치(Location) 추정과 변이(Variation) 추정

자료(Data)의 특징을 일반적으로 위치와 변이로 나타낼 수 있음

## 위치 추정 = “중심 경향성”

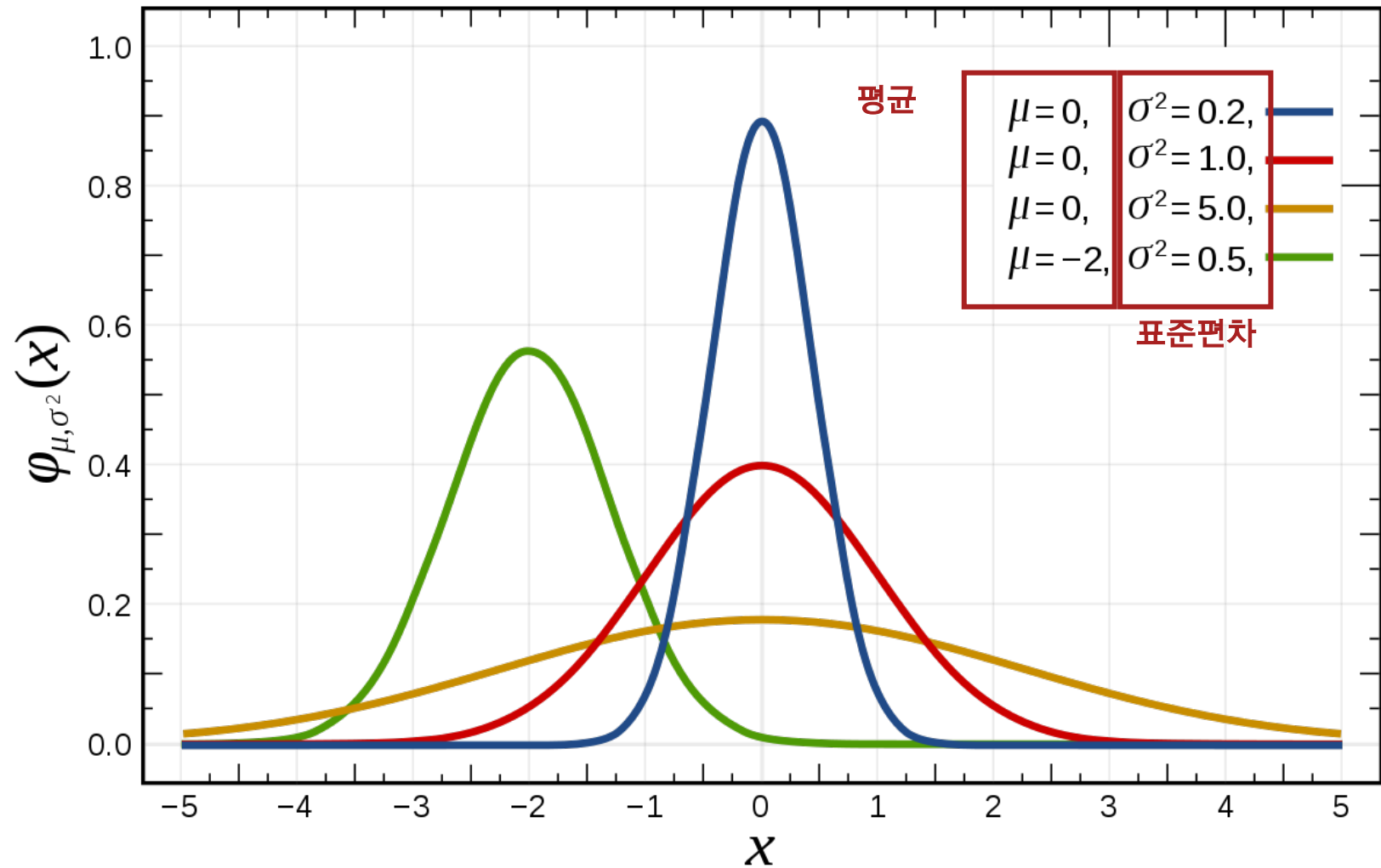
- 가장 기초적인 자료의 특징 표현 방법은 자료의 대표값 즉, ‘대략 어디쯤 위치하는구나’ 라는 중심 경향성을 나타내는 것임
- 위치를 나타내는 통계치에는 평균, 중간값, 가중 평균값, 절단 평균값 등이 있음

## 변이 추정 = “산포도”

- 데이터가 얼마나 밀집해 있는지, 얼마나 퍼져 있는지 산포도(Dispersion)을 나타냄
- 변이가 평균 근처에 몰려 있다면 통계치의 “신뢰성”이 높다고 말할 수 있음

두 정보를 종합해 계수 추정(관계의 정도), 가설검정(관계 정도의 유의성) 등을 할 수 있음

# 예를 들어 정규분포는...





# 중심경향 : 평균(Mean), 중앙값(Median), 최빈값(Mode)

## 평균 (Mean)

- 일반적으로 평균이라고 하면 산술평균을 의미함. 평균은 지나치게 크거나 작은 값에 영향을 받을 수 있음

$$\text{모평균 : } \mu = \frac{x_1 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i, \text{ 표본평균 : } \bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

## 중앙값 (Median)

- 주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미
- 지나치게 크거나 작은 값에 영향을 받지 않아 이상치를 판단할 수 있음

## 최빈값 (Mode)

- 가장 빈도가 높은 자료

# 산포 정도 : 분산(variance) 및 표준편차(Standard deviation)

## 분산 (Variance)

- 자료가 평균(Mean) 혹은 기대값(Expectation)으로부터 얼마나 떨어진 곳에 분포하고 있는지를 가늠하는 지표로 자료가 퍼져있는 정도를 나타냄
- 일반적으로 예측모형의 경우, 예측값의 분산을 최소화 하는 모형 즉, 예측값의 변동이 적은 모형을 효율적인(Efficient) 모형이라고 평가함. 하지만, 단순히 분산이 작다고 예측값의 성능이 좋은 것은 아님

$$\text{모분산 : } \sigma^2 = \frac{(x_1 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2,$$

$$\text{표본분산 : } s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

## 표준편차 (Standard Deviation)

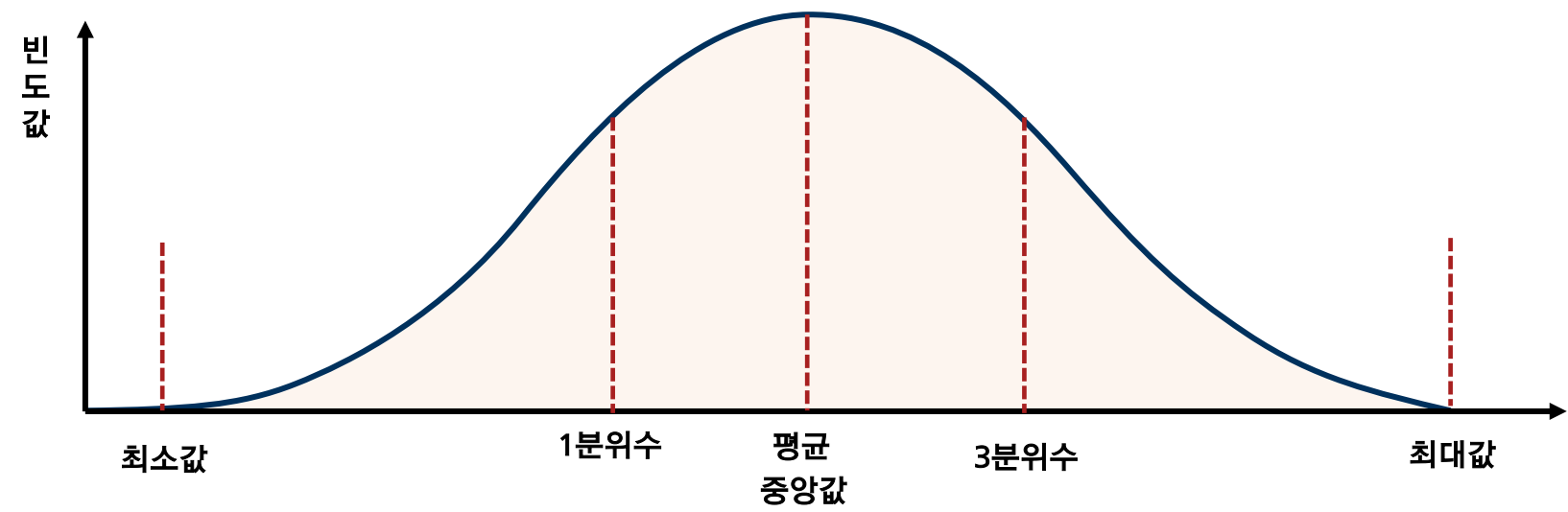
- 표준편차 역시 자료의 산포도를 나타내며 분산의 양의 제곱근으로 정의됨

$$\text{모표준편차 : } \sigma = \sqrt{\frac{(x_1 - \mu)^2 + \dots + (x_N - \mu)^2}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2},$$

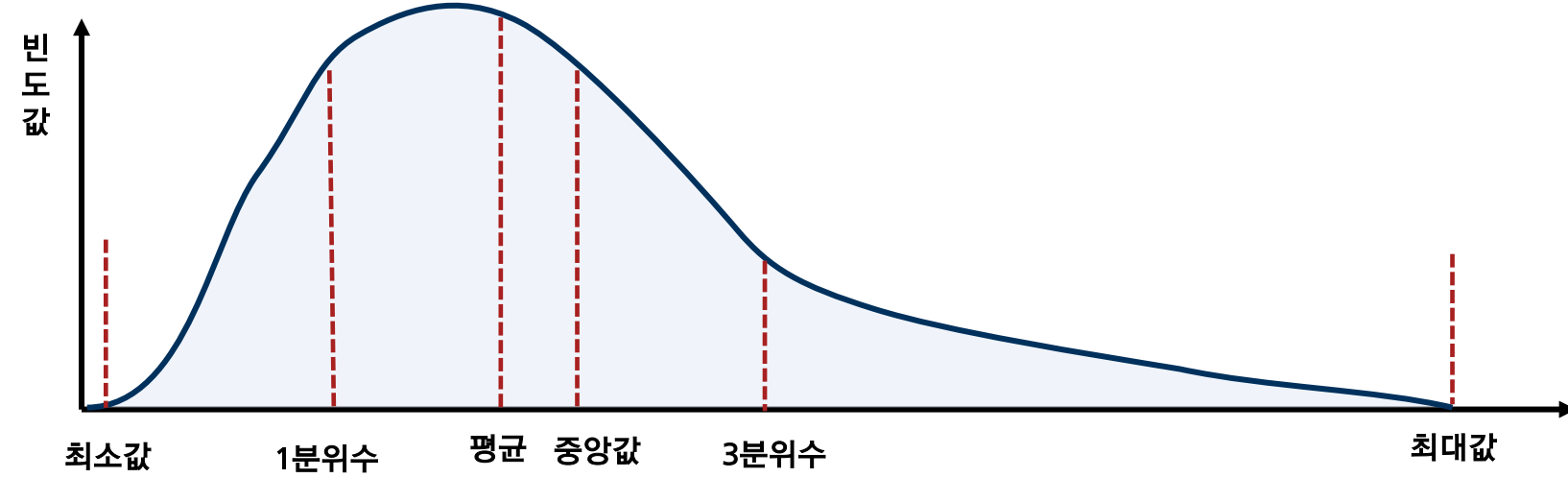
$$\text{표본표준편차(표준오차) : } s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

# 산포 정도 : 사분위수(Quartile)

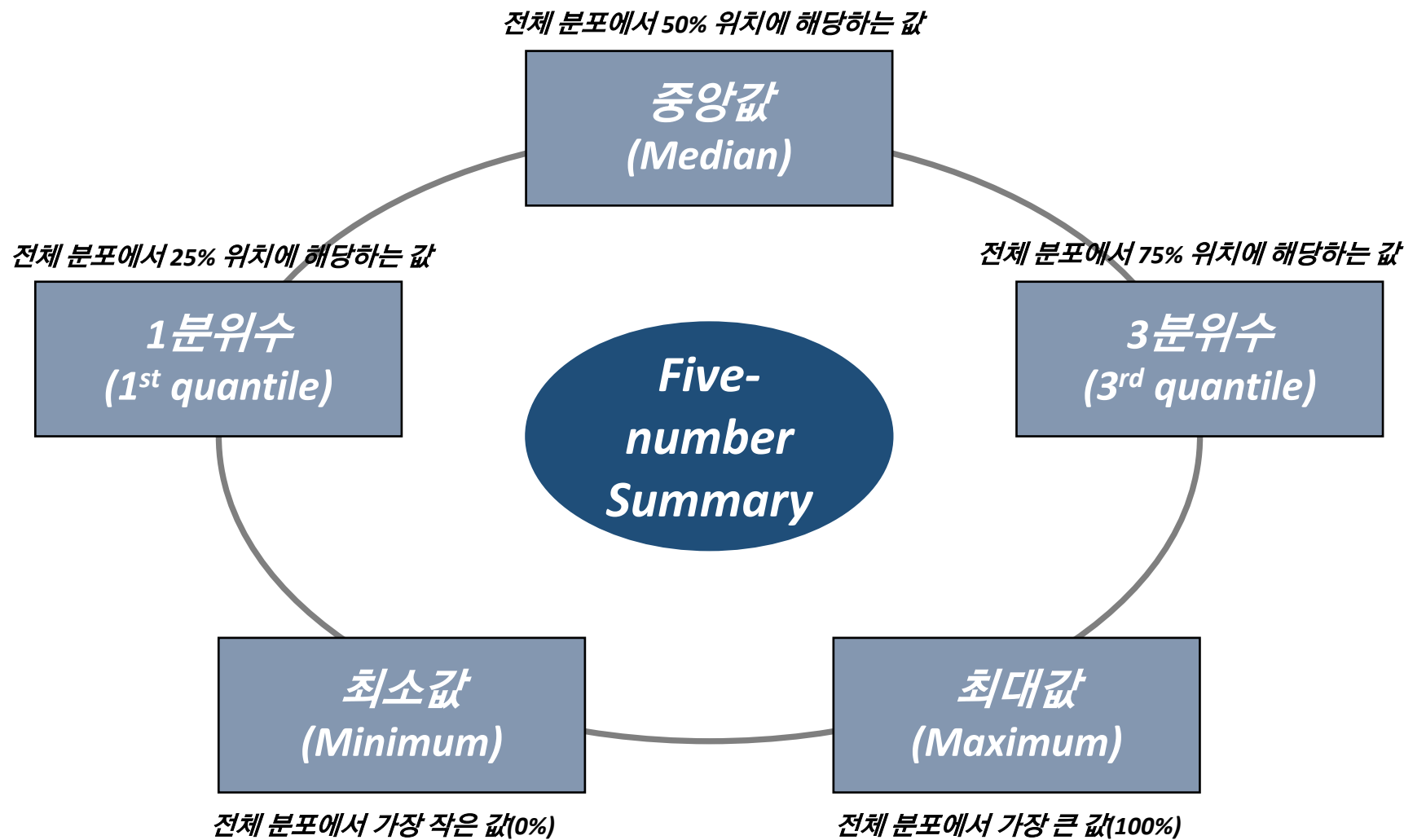
예시1



예시2



# Five-number summary



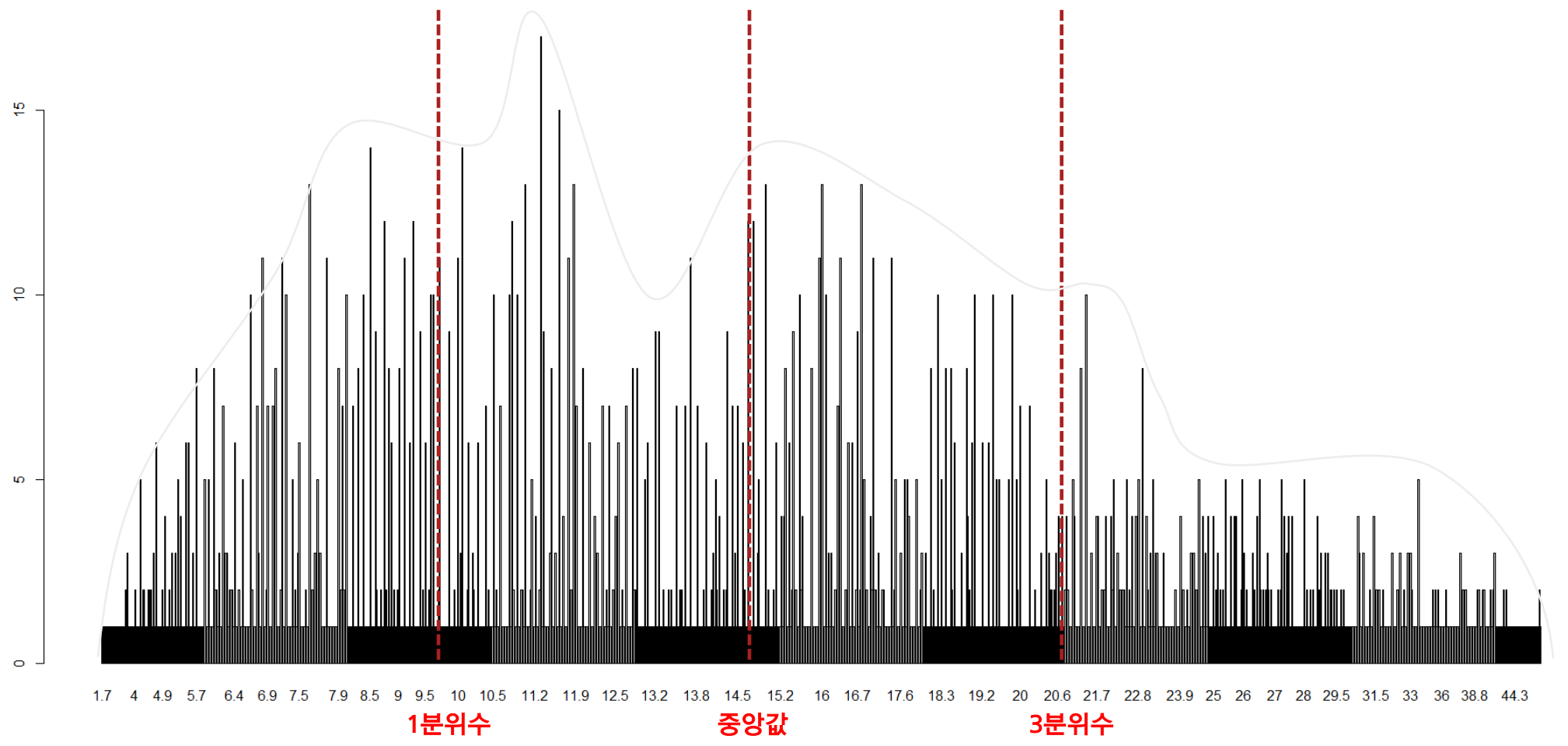
## 다시 표를 살펴보자.

Chicago의 일 평균 초미세먼지(pm2.5) (1987년 1월 1일 ~ 2005 12월 31일)

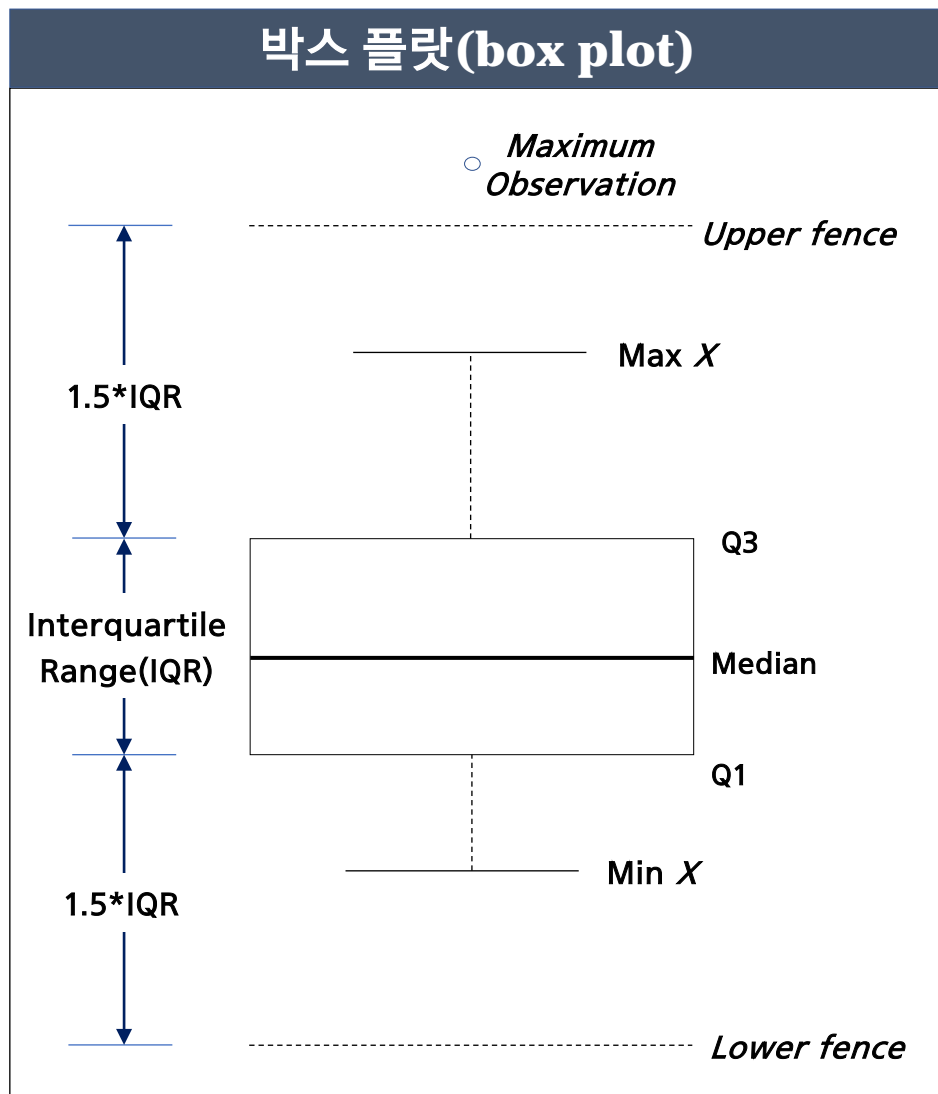


최소값 (Min)	1.70
1분위수 (1 <sup>st</sup> Quantile)	9.70
중앙값 (Median)	14.66
3분위수 (3 <sup>rd</sup> Quantile)	20.60
최대값 (Max)	61.50
평균 (Mean)	16.23
표준편차 (Standard Deviation)	8.7

# 그림으로 그려보면...



# 박스 플롯(Box plot)을 이용한 기술통계량



## *Median*

- 자료의 중위수(중앙값)를 나타냄

## *Q1, Q3*

- Q1은 사분위수 중 25%에 해당하는 1사분위수이고, Q3은 사분위수 중 75%에 해당하는 3사분위수임

## *Upper fence / Lower fence*

- Upper fence와 Lower fence는 각각 Q3, Q1의 1.5배에 해당하는 범위

## *Max X / Min X*

- Upper fence와 Lower fence 내에서 관측값의 최대값(Max)과 최소값(Min)을 나타냄

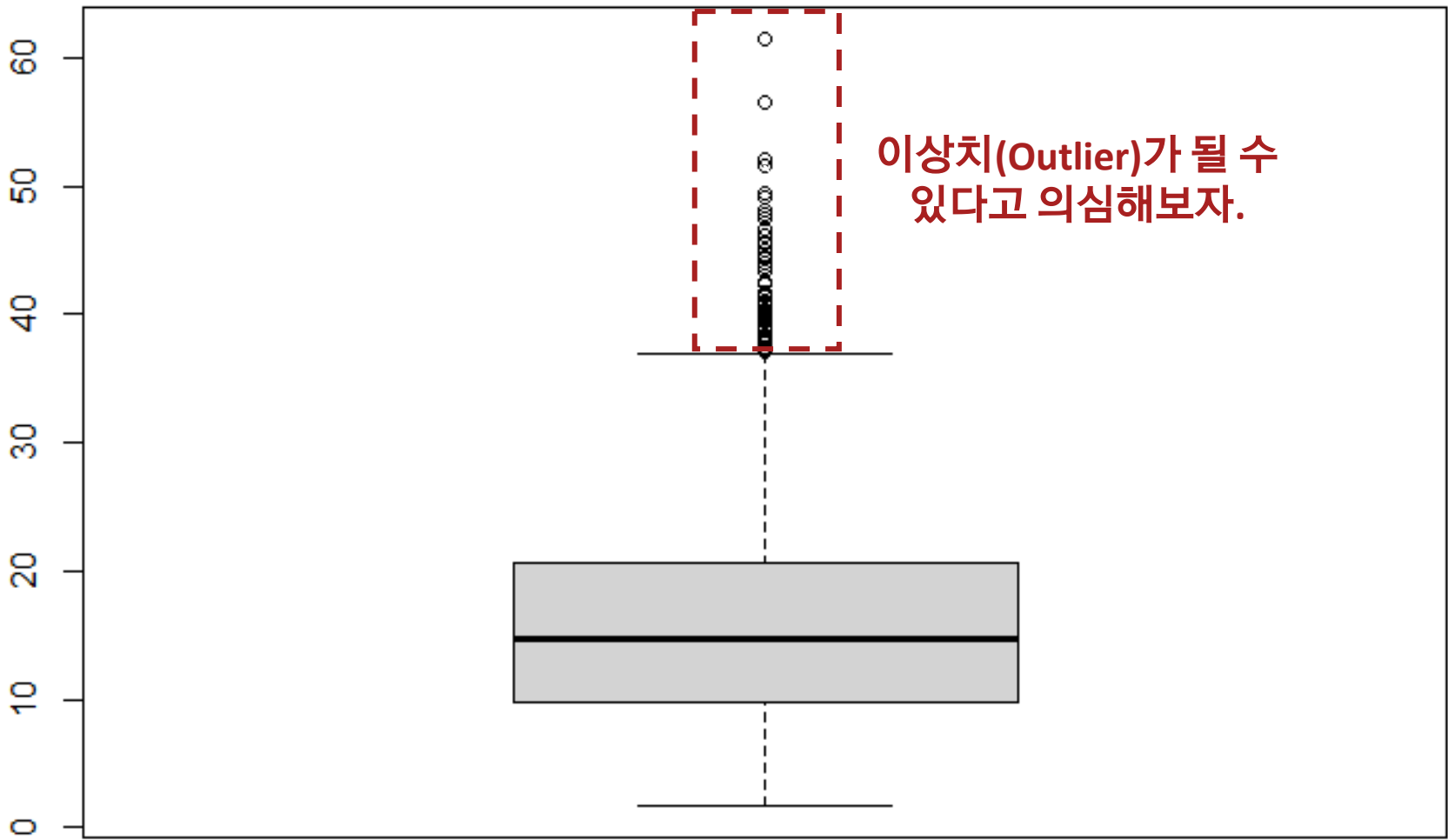
## *Maximum Observation(Outlier)*

- 전체 관측치 중 fence 밖의 최대값 혹은 최소값을 나타냄

## *Interquartile Range(IQR)*

- Q1과 Q3까지의 거리를 나타냄

# 시카고 초미세먼지 농도의 Boxplot

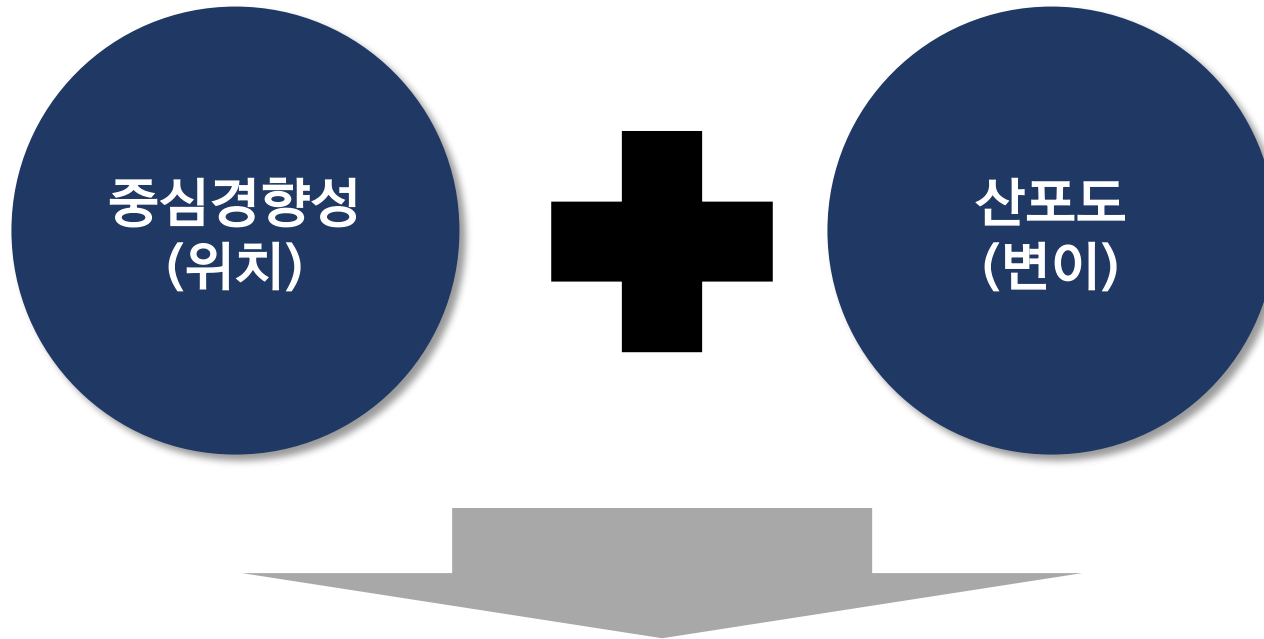




# 데이터를 바라보는 두 가지 관점

중심 경향은 어떠한가?

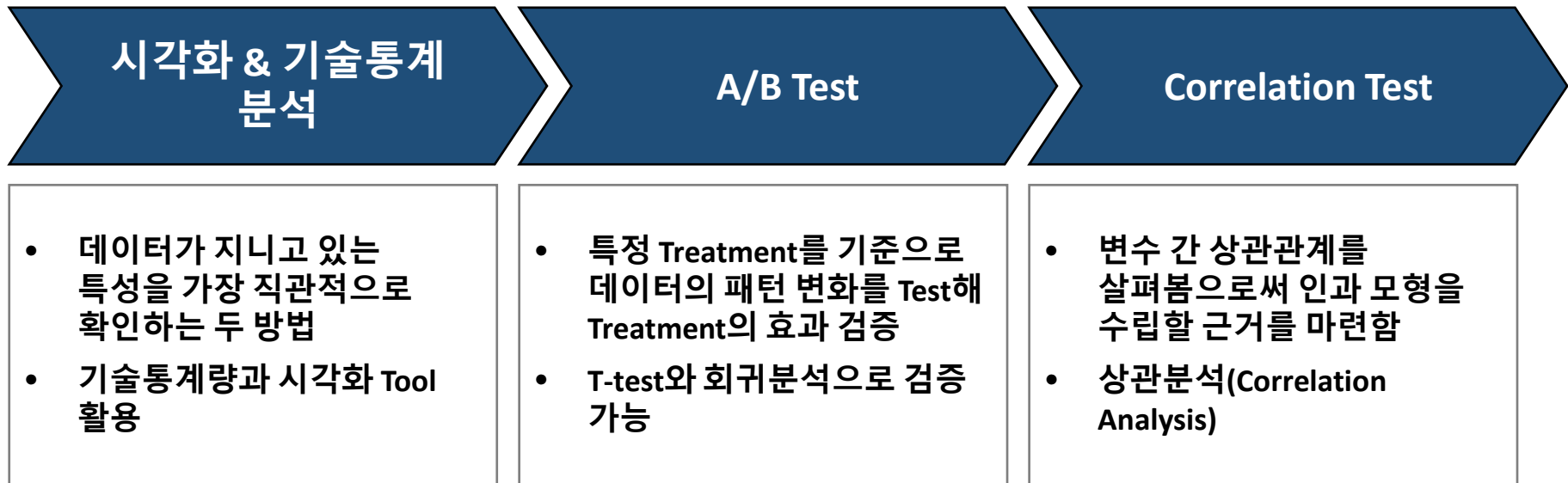
어떻게 흩어져 있는가?



주어진 데이터의 분포에 대한 “감”을 가질 수 있고,  
분석 모형 수립의 출발점이 될 수 있음

Lecture 4-2

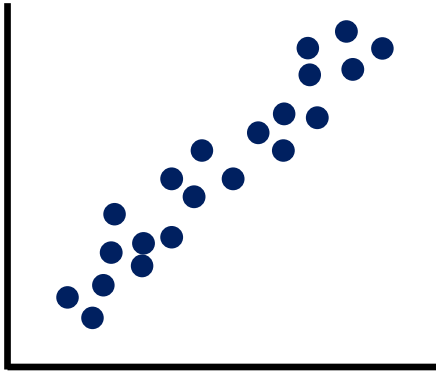
**탐색적 자료분석  
이란?**



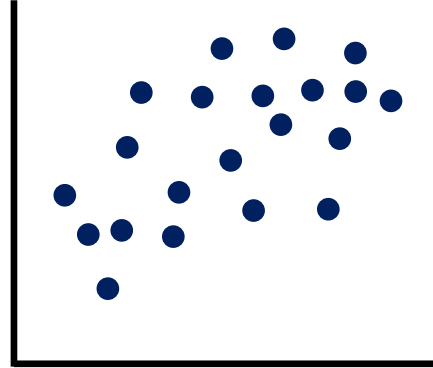
# 상관관계(Correlation)란 무엇인가?

어떤 데이터가 상관관계(Correlation)이 높다고 말할 수 있는가?

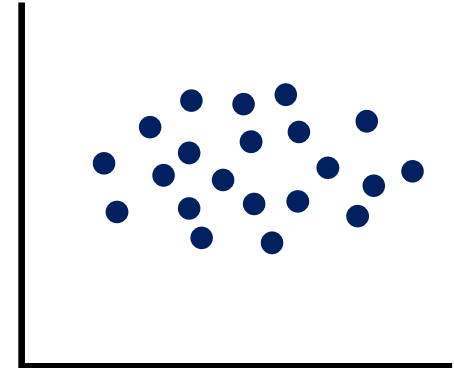
(A) 양(+)의 (강한) 상관관계



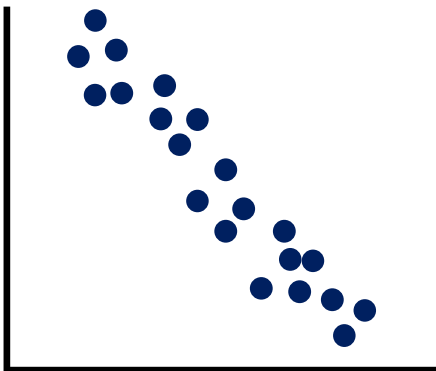
(B) 양(+)의 (약한) 상관관계



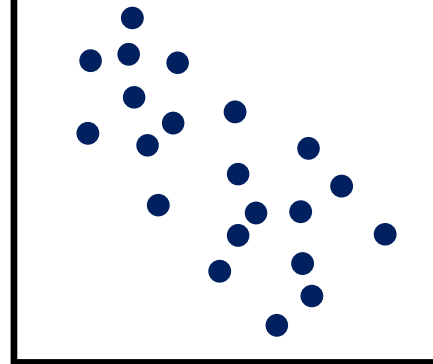
(C) 상관관계가 거의 없음



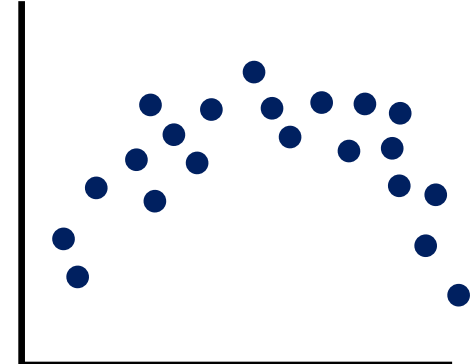
(D) 음(-)의 (강한) 상관관계



(E) 음(-)의 (약한) 상관관계



(F) 비선형 상관관계

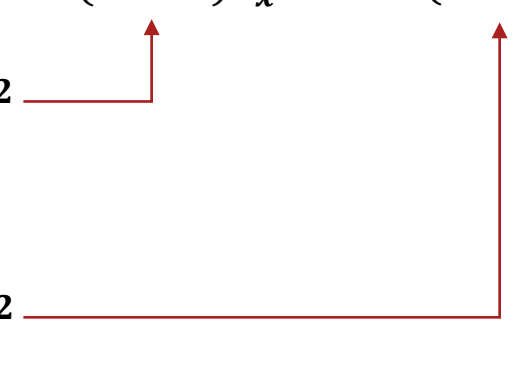


# 상관관계(Correlation)와 상관계수(Correlation Coefficient)

상관관계는 한 변량이 (선형적으로) 변화할 때, 다른 변량도 함께 (선형적으로) 변화하는 것을 말하는데 상관관계의 정도를 나타내는 값으로 피어슨(Pearson)이 개발한 피어슨의 상관계수를 많이 활용함

$$-1 \leq r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\text{공변량}}{(n-1)s_x s_y} \leq 1$$

$\quad \quad \quad = (n-1)s_x^2 \quad \quad = (n-1)s_y^2$

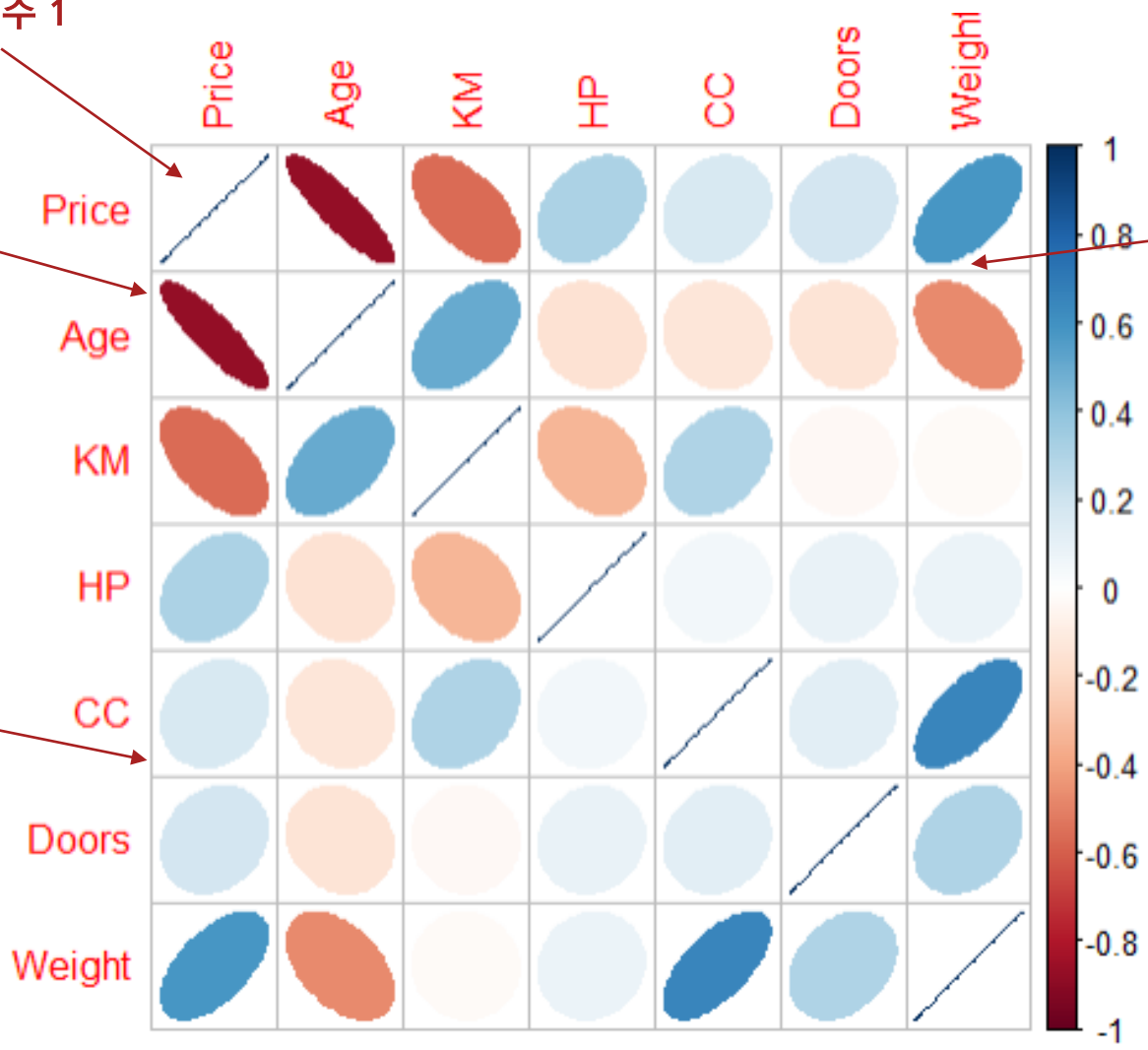
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^I (x - \bar{x})^2$$
$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^I (y - \bar{y})^2$$


# 상관관계 행렬(Correlation Matrix)

대각선은 항상 상관계수 1

상관관계가  
높을수록  
타원이 직선에  
가까워짐

상관관계가  
낮을수록  
타원이 원에  
가까워짐



상관관계의 방향(양 or 음)에 따라 원의 색(Color)과 방향(Direction)이 달라짐

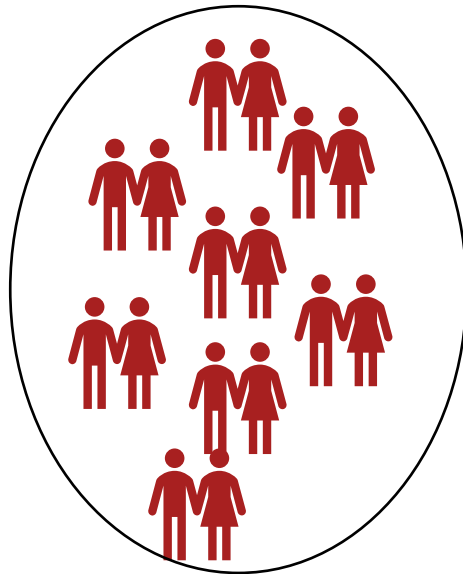
## Lecture 3-3

### 추정과 검정

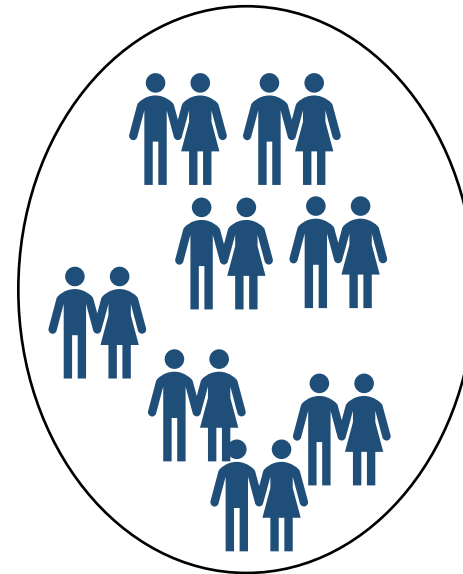
# 통계적 실험의 중요성

대부분 우리가 데이터를 통해 얻고자 하는 통계적 추론(Inference) 문제는, 처치(Treatment)에 대한 가설과 이에 대한 검정을 목표로 하는 경우가 많음. 통계적 실험 역시 일종의 인과관계 검증 과정임

프로모션 제공 집단



프로모션 미제공 집단



VS

*“ 프로모션의 효과가 있는가? ”*

이미 우리는 처치에 대한 효과를 직관적으로 추론하고 있다.  
통계적 실험은 직관적 추론을 통계적으로 검증하는 과정이라 할 수 있다.



# 가설의 수립과 검정

- $H_0$  (귀무가설, 영가설, Null hypothesis) : 기각(Reject) 하고자 하는 가설 (인과관계 때문이 아니라 우연이다 라고 주장하는 가설)

$$H_0 : \mu_A(\text{Group A 판매량 평균}) = \mu_B(\text{Group B 판매량 평균})$$

- $H_1$  (대립가설, Alternative hypothesis) : 채택(Accept) 하고자 하는 가설 (귀무가설이 기각된다면 자동으로 채택되는 가설)

$$H_1 : \mu_A(\text{Group A 판매량 평균}) \neq \mu_B(\text{Group B 판매량 평균})$$

## Q1. 왜 귀무가설을 Main으로 다루는가?

- 우리는 “효과가 있는지” 여부를 검정하는 것에 관심이 있지, “효과가 얼마나 있느냐”에 대한 문제는 훨씬 복잡함
- 만약, 귀무가설에서  $\mu_A = \mu_B$  가 기각(Reject) 된다면, 즉, ' $\mu_A \neq \mu_B$  는 우연히 몇 번 나온게 아니다' 라고 결론 내릴 수 있다면 우리가 목표로 한 처리(Treatment)의 효과는 검증되는 것이고, 그때 평균적으로 그 효과에 의해  $|\mu_A - \mu_B|$  만큼 차이난다고 인과관계를 설명할 수 있게 됨

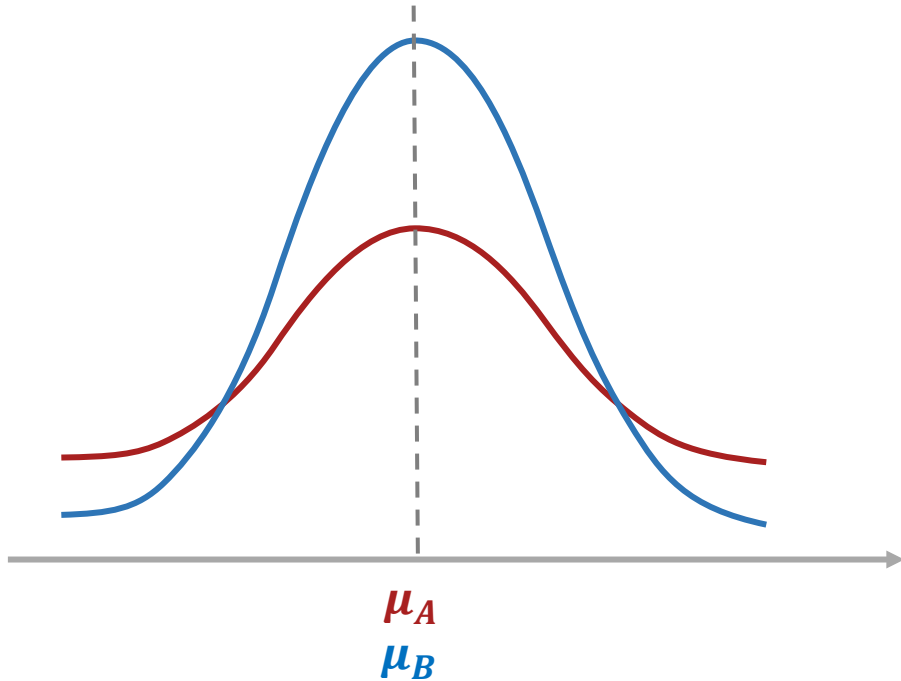
## Q2. 그렇다면, $H_0$ 를 어떻게 검정(Test)할 수 있는가?

- 대부분의 통계적 실험은 실험군과 대조군의 분포(Distribution)를 비교함
- 두 집단의 평균 차이가 실제 차이에 의한 것인지, 우연에 의한 것인지 1) '재표본추출'을 통해 휴리스틱(Heuristic)하게 검증하는 방법과 2) T-분포를 이용해 통계적으로 검정하는 두 가지 방법이 있음

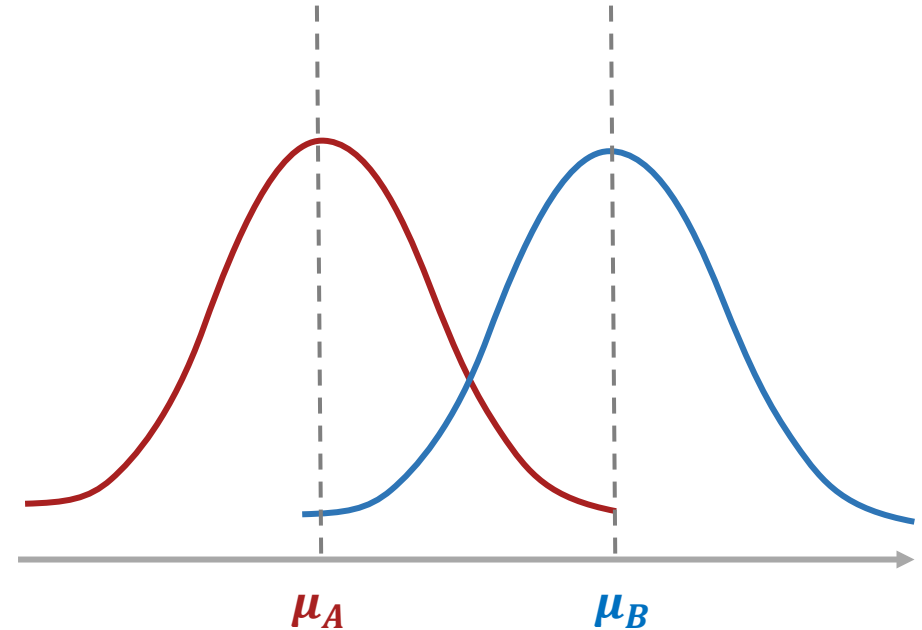
# A/B 테스트에서 두 집단의 비교 (1)

두 집단의 차이가 있다고 말할 수 있는 Case는 어떤 경우일까?

Case 1)



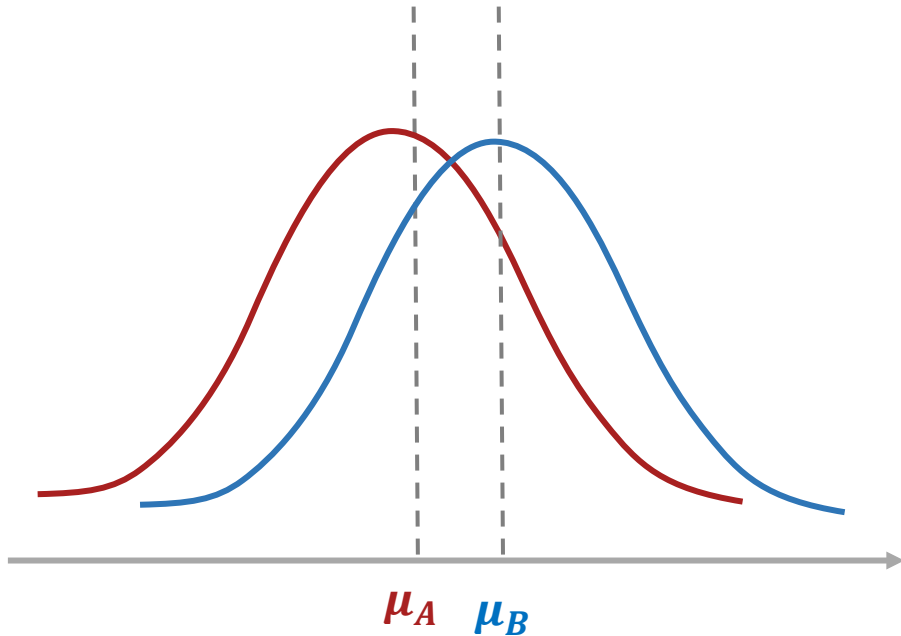
Case 2)



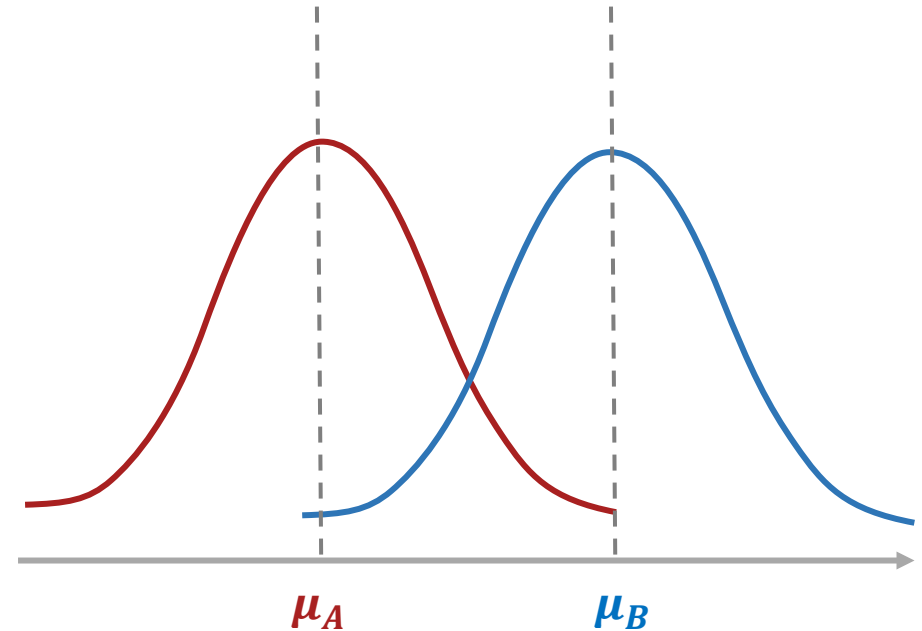
# A/B테스트에서 두 집단의 비교 (2)

두 집단의 차이가 있다고 말할 수 있는 Case는 어떤 경우일까?

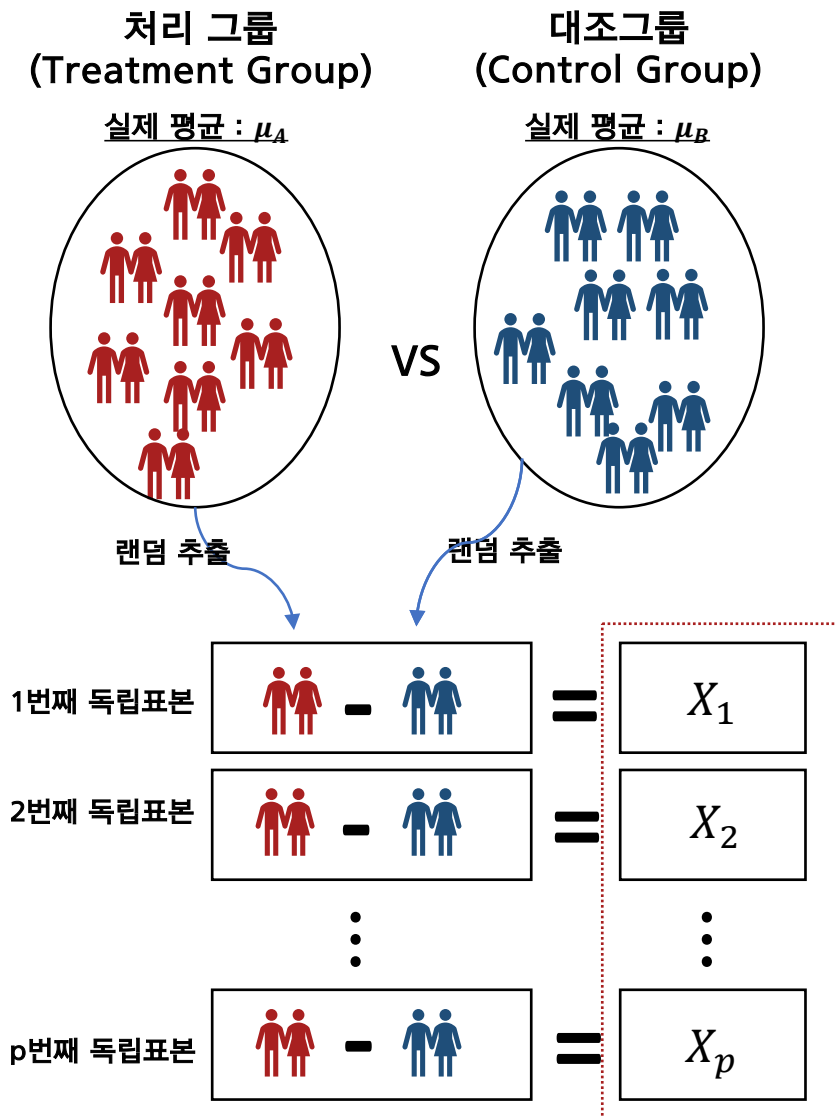
Case 1)



Case 2)

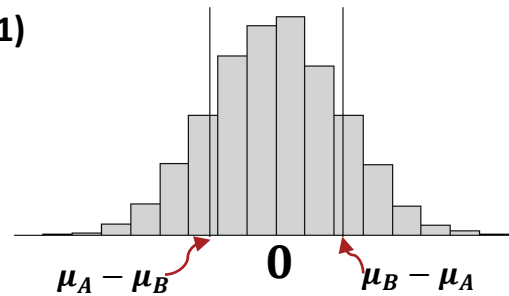


# 두 집단의 차이 분석

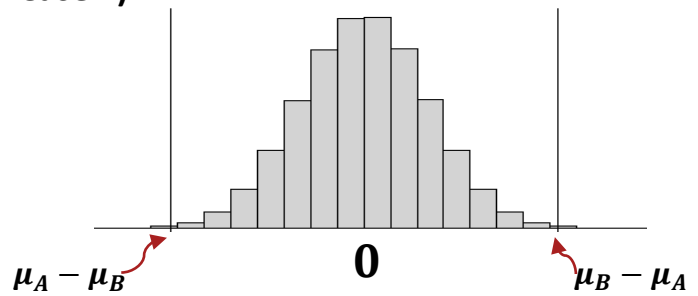


$X_1, X_2, \dots, X_p$ 들의 분포를 나타내면 다음과 같다고 하자.

Case 1)

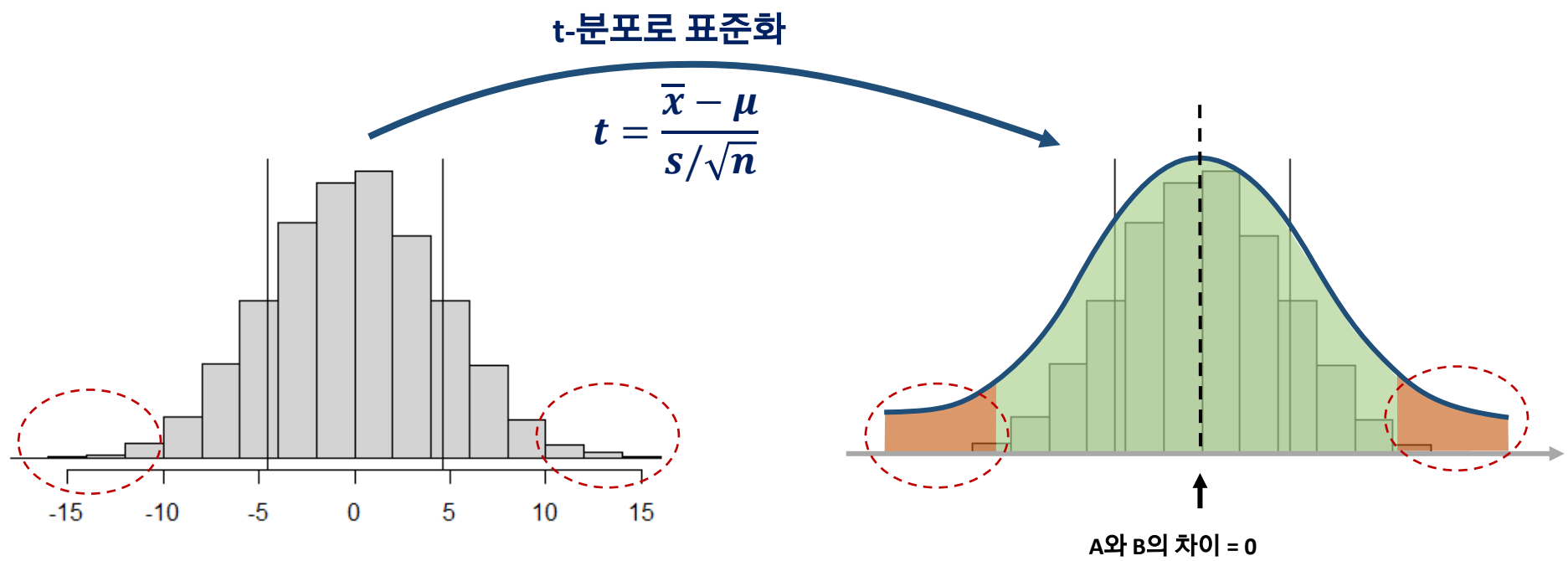


Case 2)



# T-Test (T-검정)

통계적으로 가설을 검정한다는 것은 “두 현상”이 나타내는 통계분포가 같은지(귀무가설), 다른지(대립가설)를 검정하는 것인데, 일반적으로 t-분포가 같은가 여부로 비교함



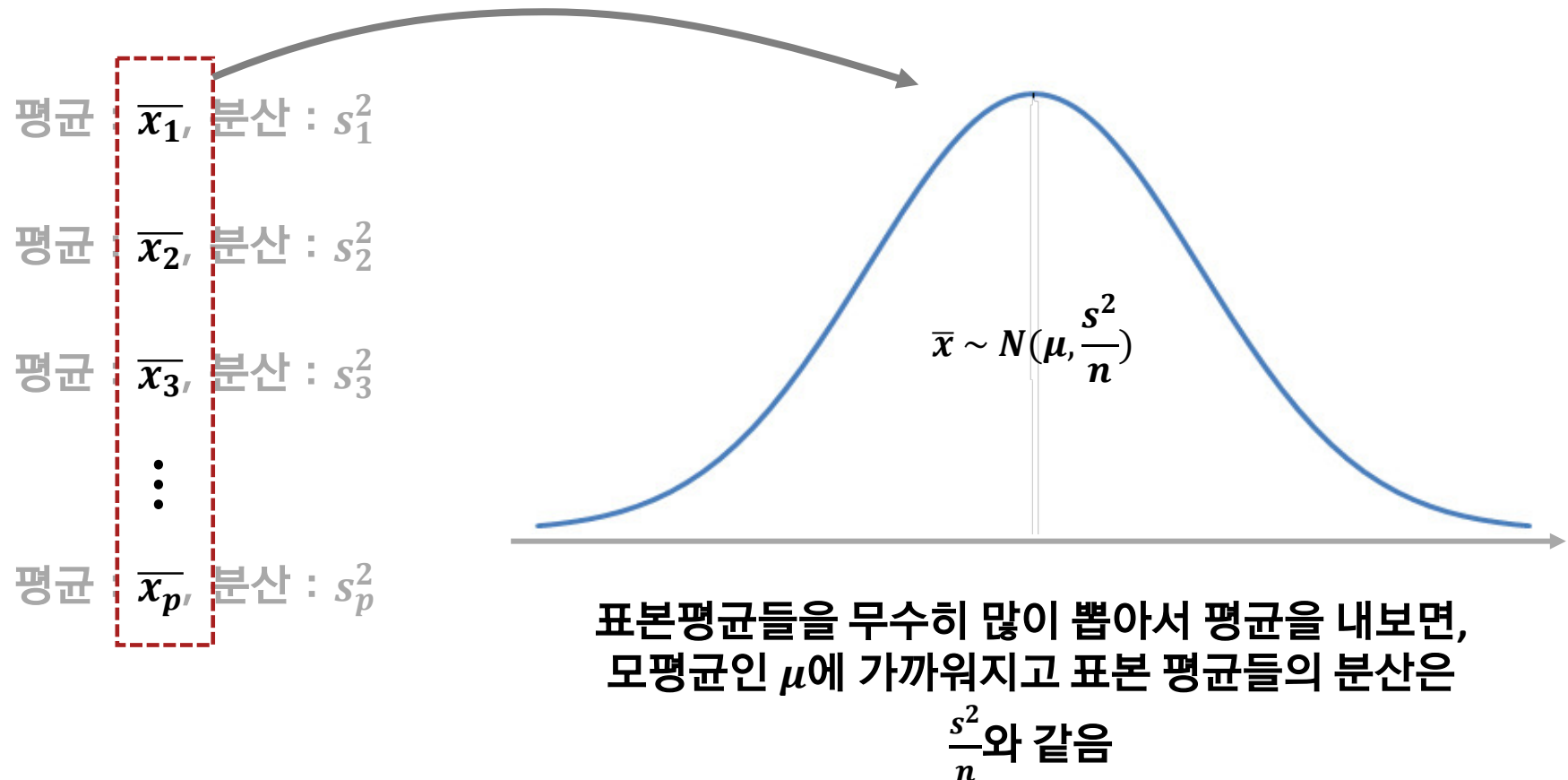
실제 차이가 있다라고 말할 수 있으려면, 즉,  
통계적으로 차이가 있다고 말할 수 있으려면  
차이의 평균이 (                      )에 있어야 한다 !

=

위 그림 상에서 (노란색/주황색) 영역에 있어야  
한다.

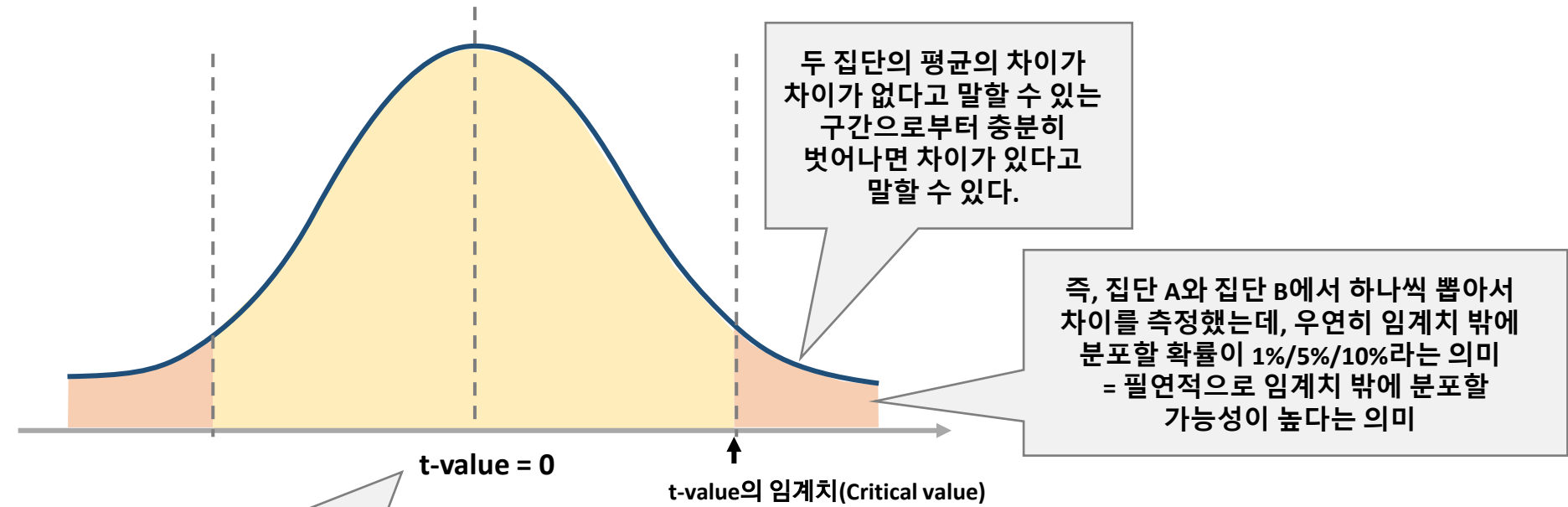
# 참고 - 왜 t-분포가 같은지를 검증하는가?

표본평균을 무한히 생성하면, 모평균이 어떤 분포를 따르든 상관없이 중심극한정리에 의해 정규성(Normality)을 따른다고 할 수 있음



# T-분포에서 t-value와 p-value의 개념 이해

## 집단 A와 집단 B의 평균 차이의 t-분포



두 집단 간 평균차이가  
( 있다 / 없다 )

$\alpha$ -수준에서 통계적으로 유의하려면	즉, p-value가	그러려면, t-value의 절대값은
10%	0.01보다 작아야 함	약 1.65 보다 커야 함
5%	0.05보다 작아야 함	약 1.98 보다 커야 함
1%	0.01보다 작아야 함	약 2.56보다 커야 함

# 통계적 유의성과 p 값

검정(Testing)을 이해하는 데 있어서 가장 중요한 개념이 “통계적 유의성”과 “p-value(p값)”임.  
 통계적 유의성은 관측된 결과가 우연인지, 실제 영향이 있는 것인지를 판단하는 것을 의미함

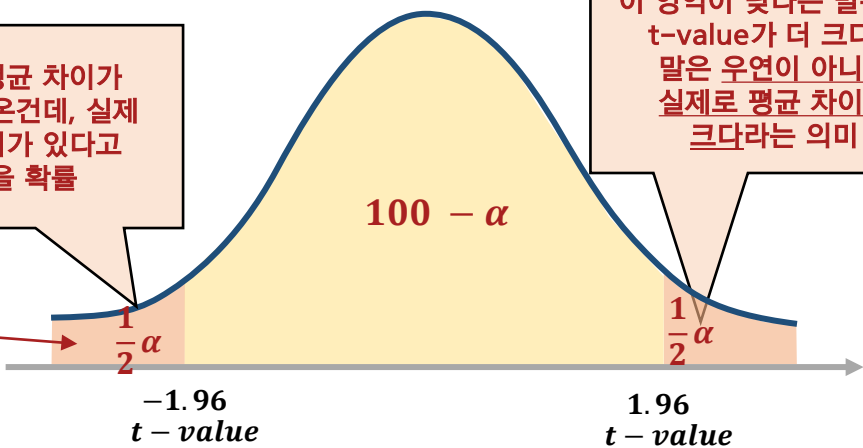
※ 제1종 오류(유의수준)와 제2종 오류

		실제 결과	
		실제로 효과가 있음	우연히 효과 관측됨
의사결정	귀무가설 기각	True	$\alpha$
	귀무가설 채택	$\beta$	True

“ Type I Error ” : 우연히 관측된 결과인데, 실제 효과가 있다라고 예측할 확률

우연히 평균 차이가 있다고 나온건데, 실제 평균 차이가 있다고 나왔을 확률

이 영역이 낮다는 말은 즉, t-value가 더 크다는 말은 우연이 아니라 실제로 평균 차이가 크다는 의미



“ Type II Error ” : 실제 효과가 있는 건데, 우연히 관측된 결과라고 기각할 확률

- ✓ 이때, p-value가 0.05보다 작으면, 즉, 5%보다 작으면 “5% 유의수준에서 평균의 차이가 유의하다.” 라고 말할 수 있음
- ✓ P-value가 0.01보다 작으면, “1% 유의수준에서 평균의 차이가 유의하다(통계적 유의성이 더 높다.)”



# 신뢰구간(Confidence Interval)

## ➤ 모평균 구간 추정 (모평균 $\mu$ 의 $100(1-\alpha)\%$ 신뢰구간)

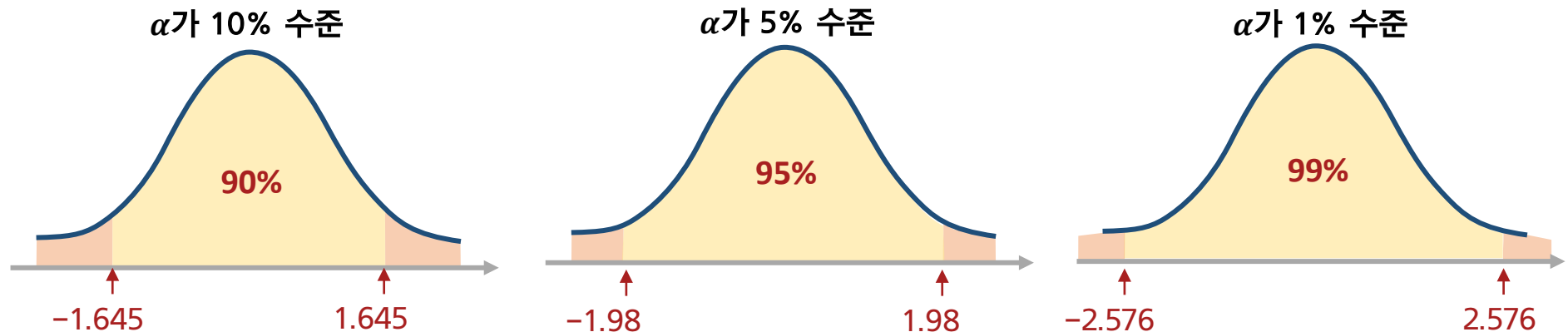
- Z분포의 유의수준  $\alpha$ 에서 신뢰구간 :

$$P\left(-Z_{\frac{\alpha}{2}} \leq \frac{x-\mu}{\sigma/\sqrt{n}} \leq Z_{\frac{\alpha}{2}}\right) = P\left(\mu - Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq x \leq \mu + Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- T분포의 유의수준  $\alpha$ 에서 신뢰구간 :

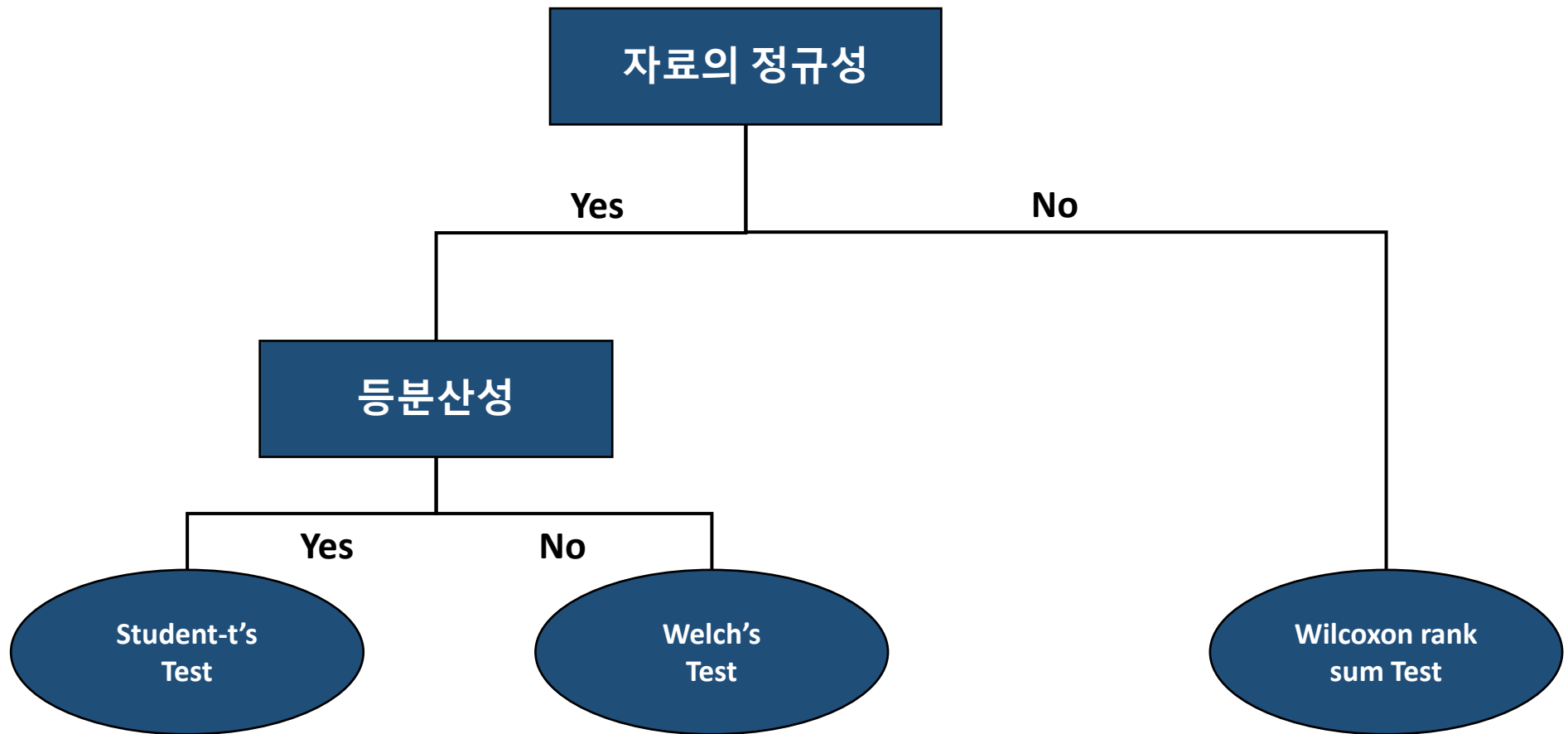
$$P\left(-t_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t_{\frac{\alpha}{2}}\right) = P\left(\mu - t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq \bar{x} \leq \mu + t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

- 만약, 표본 수가 충분히 크다면  $t_{\frac{\alpha}{2}}$  값은 아래와 같음



# T-test 를 위한 기본 가정

T-test는 자료의 정규성 및 동분산성에 따라 검정방법을 구분할 수 있음



# T-test의 유형

t-test는 비교 표본에 따라 3가지 유형으로 구분할 수 있음

## 독립표본 t-test (Independent two sample t-test)

- 서로 다른 두 집단 간 평균 차이를 비교할 때 활용
- 두 표본이 서로 관계없는 모집단에서 추출되어야 하고, 표본 간 아무런 관계가 없어야 함

Ex) 프로모션을 적용한 소비자 집단과 그렇지 않은 집단 간 구매 빈도 비교

## 대응표본 t-test (Paired sample t-test)

- 동일 집단에 대해 처리 전/후 차이를 비교할 때 활용
- 대응 표본은 비교대상이 1:1 매칭되어야 하므로 샘플 수가 같아야 함

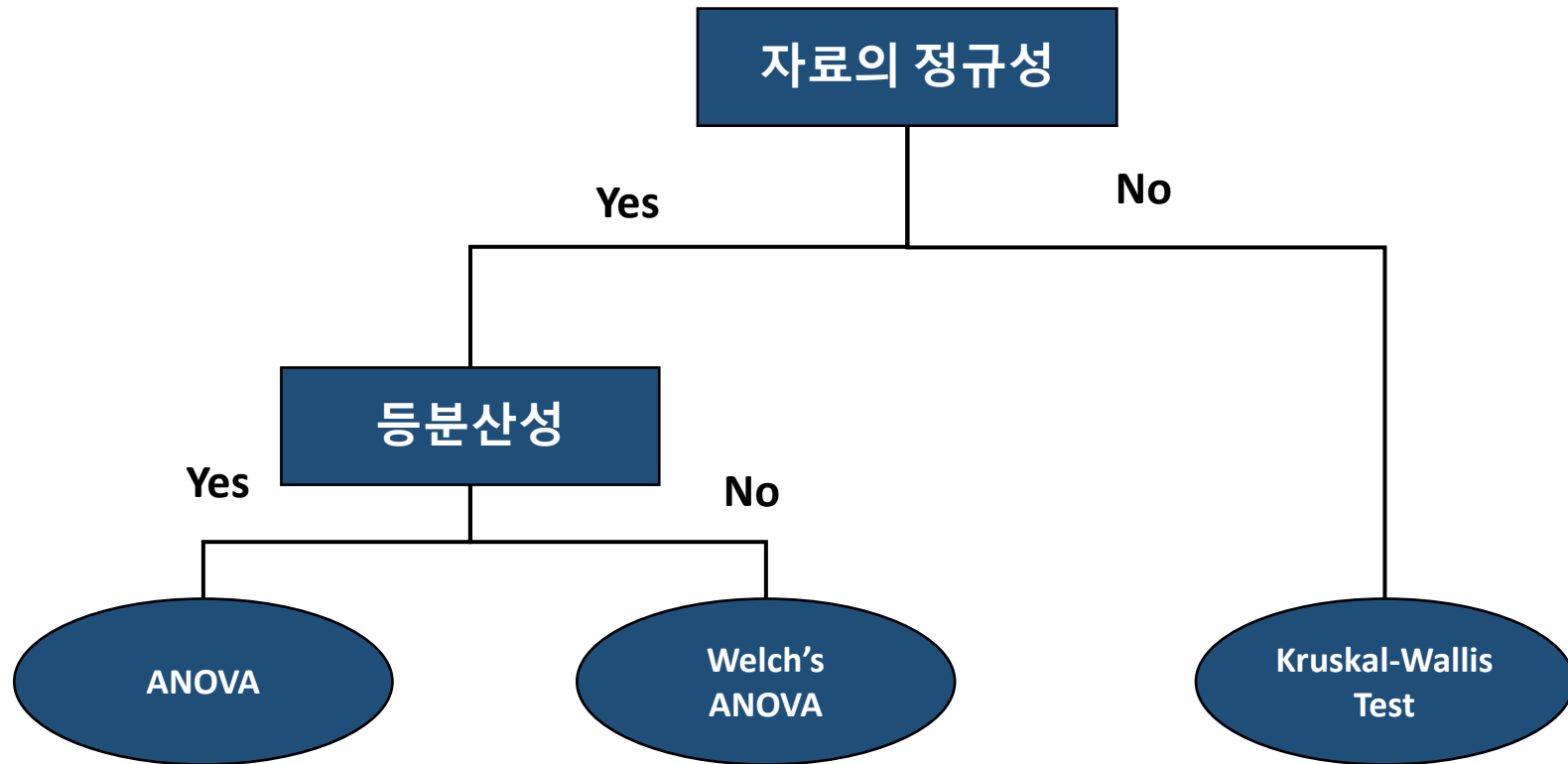
Ex) 30명의 환자들에게 신약 투약 전/후 암세포 크기 변화 비교

## 일표본 t-test (One sample t-test)

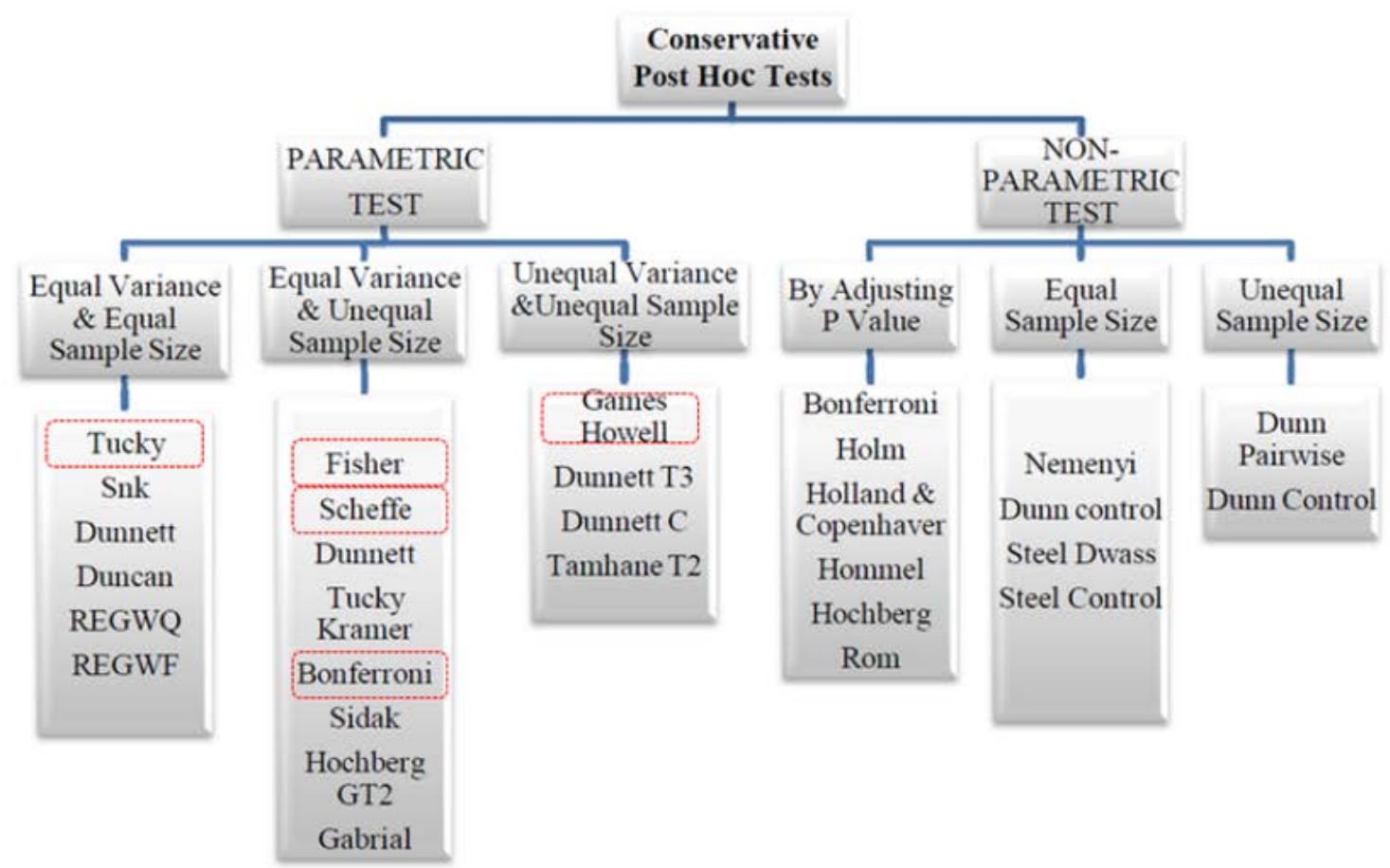
- 하나의 집단의 평균이 특정 값(연구자가 정한 값)과 차이가 있는지 혹은 큰지/작은지 비교할 때 활용

Ex) 중간고사 평균이 78점인 학급이 기준치인 75점보다 유의하게 높다고 할 수 있는지 비교

# 만약, 3개 이상의 Group에 대한 Test를 하려면 ?



# ANOVA와 Kruskal-Wallis 검정을 위한 사후검정 방법



Shingala, M. C., & Rajyaguru, A. (2015). Comparison of post hoc tests for unequal variance. International Journal of New Technologies in Science and Engineering, 2(5), 22-33.