

# 가설검정 및 유의확률 T-test

6조

# 1

가설검정  
P-value  
T-test

신해솔

# What is Hypothesis?

어떤 현상은 이럴 것이다”라고  
잠정적 진술을 하고,  
이에 대해 옳고 그름을 판단하는 의사결정을  
하게 되는데,  
이 때 잠정적 진술을 가설이라고 한다.

영가설( $H_0$ )

VS.

대립가설( $H_A$ )

기각 될 것을 기대하고  
수립되는 가설 (= 귀무가설)

일반적으로 연구자가  
주장하고자 하는 가설  
(= 연구가설)

영가설( $H_0$ )  
VS.  $\beta$   
대립가설( $H_A$ )

의사 결정 (통계 적 판 단)	진리 (truth)		
		$H_0$	$H_A$
	$H_0$	$1-\alpha$	(제2종 오류)
	$H_A$	(제1종 오류)	$1-\beta$ (검정 력, Po wer)

## 가설 검정 과정

1.

영가설, 대립가설 설정

2.

P-value &  $\alpha$  교

Or T-test 실시

\* Under  $H_0$

3.

결론 ( $H_0$  기각?)

What is P-value?

귀무가설이 참이라는 전제하에  
얻은 통계량이 귀무가설을  
얼마나 지지하는지를 나타낸 확률

\* '확률'의 지표점

What is  $\alpha$  ?

(= 유의도, 유의수준)

↓  
틀릴 가능성이 있는

P-Value &  
유의수준

유의수준  
파악

P-value < 유의수준

=> 귀무가설 reject

대부분  
0.01, 0.05, 0.1 사용  
-> 연구자가 정함



# P-Value 파악

```
> summary(lm(cars$speed~cars$dist))

Call:
lm(formula = cars$speed ~ cars$dist)

Residuals:
    Min       1Q   Median       3Q      Max
-7.5293 -2.1550  0.3615  2.4377  6.4179

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.28391     0.87438   9.474 1.44e-12 ***
cars$dist     0.16557     0.01749   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.156 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12
```

=> 매우 작은 값 (=0.000000000000000149)  
-> 귀무가설 reject = “car\$dist” 매우 의미(o)

# What is T-test

t검정(t test)은  
모집단의 분산(혹은 표준편차)  
을 알지 못할 때  
추정된 분산(혹은 표준편차)을  
가지고 검정하는 방법

추정치

$$\hat{\sigma}_x = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

# T-test 관련 일화

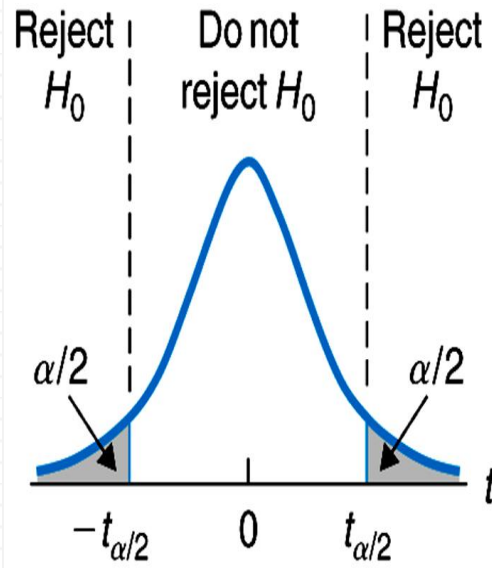


**W.S.Gosset**

-> T 분포 처음 제시

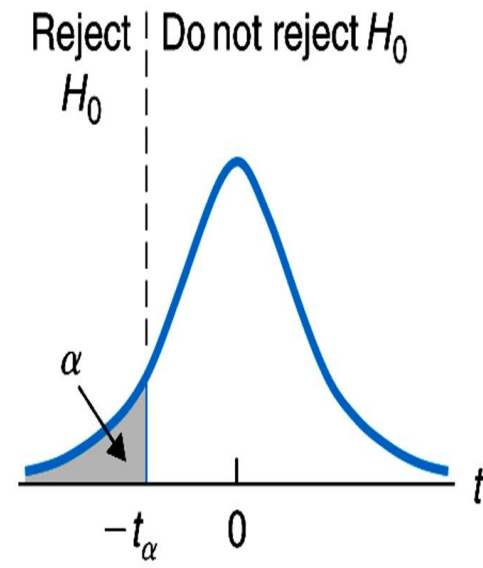
논문에서 필명으로  
Student를 사용  
오늘날까지 Student's t  
분포라고 불린다.

## T-test

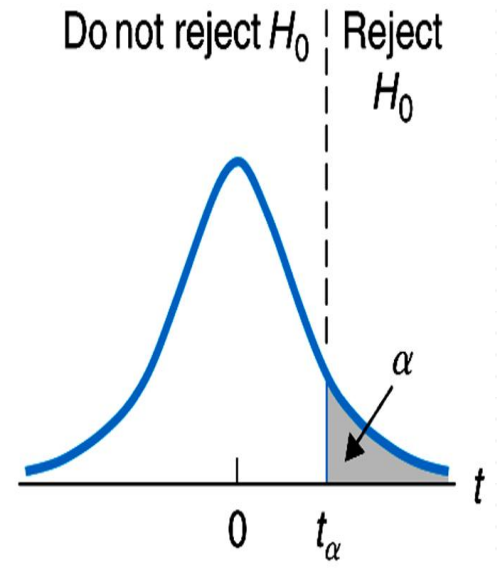


Two-tailed

양측검정



Left-tailed



Right-tailed

단측검정



T-검정통계량 > 임계값  
 $\Rightarrow$  귀무가설 기각

# 2

**Supervised Learning  
&  
Unsupervised Learning**

### Supervised learning & Unsupervised learning

#### Supervised learning

- Labeled training dataset  $D = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , where each  $x_i = (x_1, \dots, x_d)$  is associated with a label  $y_i$ .
- Input variables:  $x_1, \dots, x_d$  (d features), output variable:  $y$  (label)

#### Unsupervised learning

- Unlabeled training dataset  $D = \{x_1, x_2, \dots, x_n\}$ , where each data point  $x_i \in R^d$  contains d features  $x_i = (x_{i1}, \dots, x_{id})$
- To find useful properties/patterns of the structures of the dataset

## Supervised Learning

Input						Output
id	$X_1$	$X_2$	$X_3$	...	$X_d$	$Y$
1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1d}$	$y_1$
2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2d}$	$y_2$
3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3d}$	$y_3$
4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4d}$	$y_4$
5	$x_{51}$	$x_{52}$	$x_{53}$	...	$x_{5d}$	$y_5$
6	$x_{61}$	$x_{62}$	$x_{63}$	...	$x_{6d}$	$y_6$
7	$x_{71}$	$x_{72}$	$x_{73}$	...	$x_{7d}$	$y_7$
...	...	...	...	...	...	...

## Unsupervised Learning

Input						Output
id	$X_1$	$X_2$	$X_3$	...	$X_d$	$X$
1	$x_{11}$	$x_{12}$	$x_{13}$	...	$x_{1d}$	
2	$x_{21}$	$x_{22}$	$x_{23}$	...	$x_{2d}$	
3	$x_{31}$	$x_{32}$	$x_{33}$	...	$x_{3d}$	
4	$x_{41}$	$x_{42}$	$x_{43}$	...	$x_{4d}$	
5	$x_{51}$	$x_{52}$	$x_{53}$	...	$x_{5d}$	
6	$x_{61}$	$x_{62}$	$x_{63}$	...	$x_{6d}$	
7	$x_{71}$	$x_{72}$	$x_{73}$	...	$x_{7d}$	
...	...	...	...	...	...	

## &lt;각 학습 별 용도&gt;

**Supervised Learning**

1. Classification
  - Binary Classification
  - Multi-class Classification
  - Multi-label Classification
2. Regression

**Unsupervised Learning**

1. Data Preprocessing and Scaling
2. Dimensionality Reduction  
& Visualization
3. Clustering



## <각 학습 별 다양한 모델들>

### Supervised Learning

1. K – Nearest Neighbors
2. Linear Models
3. Decision Trees
4. Random Forests
5. Support Vector Machines
6. Neural Networks

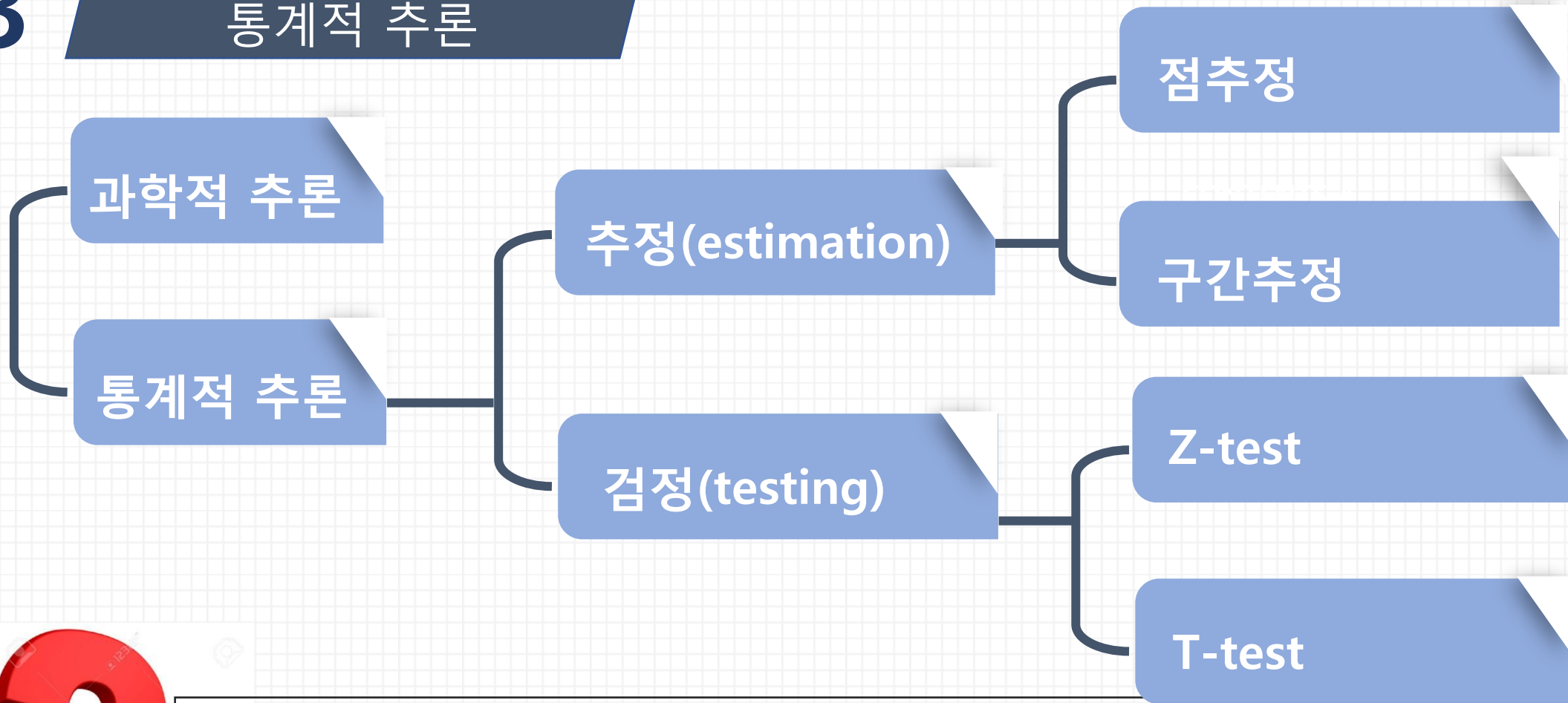
### Unsupervised Learning

1. Principal Component Analysis (PCA)
2. t-distributed Stochastic Neighbor  
Embedding (t – SNE)
3. K – means clustering
4. Hierarchical Clustering
5. DBSCAN

# 3

## T-test

김진형



- ▶ T분포 vs T-test = 신뢰구간(by t분포) vs T-test
- ★모집단 추론(estimation) vs 모집단 추정(testing)
- ▶ T-test는 추정? 검정? 검정(Testing)!

# What is T-test?

▶ (1)"두 집단 간"의 (2)"차이"가 (3)"유의미"한 지 검증하는 방법

== (1)"두 집단 간"의 (2)"평균의 차이"가 (3)"유의미"한 지

== (1)"두 모집단 간"의 (2)"평균의 차이"가 (3)"유의미"한 지

== (1)"두 집단 간"의 (2)"평균의 차이"가 (3)"존재"한다고

(4)"어느정도로 확실하게" 말할 수 있는 지

What is  
T-test?

VS

What is  
Z-test?

“표본”집단 표준편차

$s$

“모”집단 표준편차

$\sigma$

# 03 T-test를 사용하는 이유

## Z-test vs T-test

1) T-test vs Z-test

Z-test = 모집단 표준편차를 "**알고**" 있는 경우

T-test = 모집단 표준편차를 "**모르고**" 있는 경우

\*일반적으로, 우리는 모집단의 표준편차를 알지 못한다.

->따라서 T-test 를 사용한다.

## T-test를 위한 3가지 가정

1)표본이 독립적인가?

2)표본의 데이터가 정규분포를 따르는가?

3)집단이 2개인가?

# 03 T-test를 사용하는 이유

## T-test를 위한 3가지 가정

1) 표본이 독립적인가?

-> A관측치가 B표본에 의해 영향 받지 않는 것.  
Ex) A,B 상대방에 대한 호감도 조사.

<독립적>

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

2) 표본의 데이터가 정규분포를 따르는가?

-> CLT(Central Limit Theorem) 중심극한정리

(1) R의 qqnorm() + qqline()

(2) R의 shapiro.test()

3) 집단이 2개인가?

# 03 T-test를 사용하는 이유

## R\_정규분포 여부 확인

```
> #n=10  
> ex1=sample(1:30,10)  
> qqnorm(ex1); qqline(ex2)
```

```
> #n=30  
> ex2=sample(1:30,30)  
> qqnorm(ex2); qqline(ex2)
```

```
> #항목 수가 30개 미만인 벡터 데이터 생성  
> shapiro_test_vector=c(74,87,89,98,65,82,70,70,70)  
> #Shapiro-wilk 검정  
> shapiro.test(shapiro_test_vector)
```

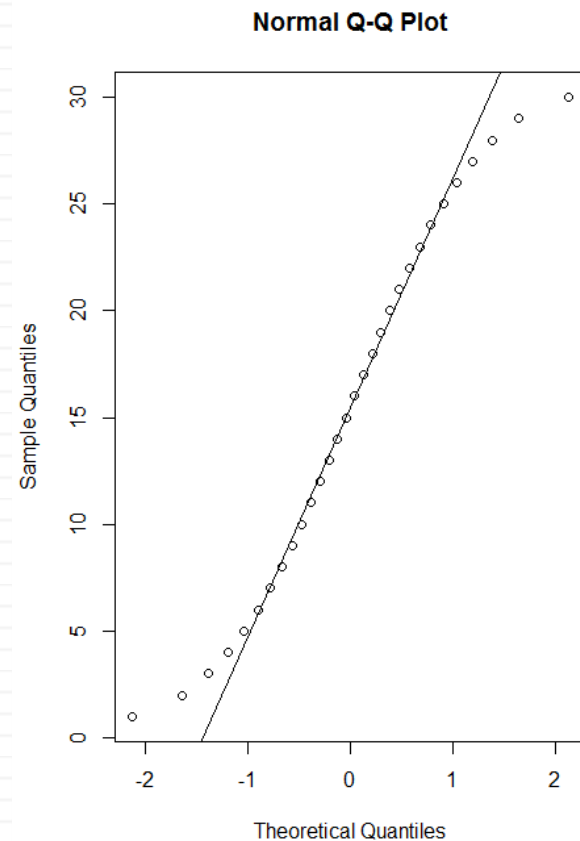
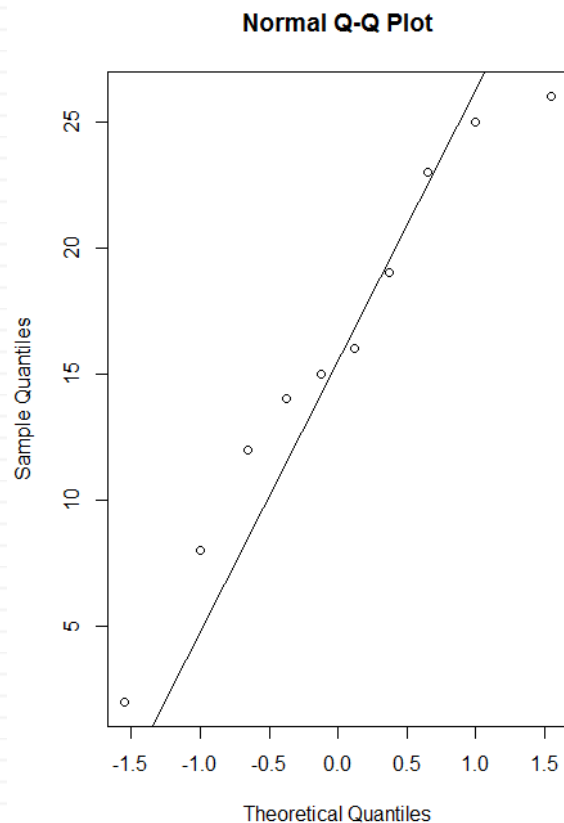
shapiro-wilk normality test

```
data: shapiro_test_vector  
W = 0.91132, p-value = 0.3252
```

$H_0$  = 정규분포를 따른다.

$H_a$  = 정규분포를 따르지 않는다.

=  $P\text{-value} > 0.05$  =  $H_0$  기각 **X** = 정규분포를 따른다





# 03 T-test

## T-test

▶ `t.test(x,y,paired=TRUE,var.equal=TRUE,alternative="two.sided")`

입력 항목	설명
X	A집단 벡터
Y	B집단 벡터
Paired	대응표본인 경우 TRUE, 독립표본인 경우 FALSE
Var.equal	두 집단의 분산이 같다면 TRUE, 다르면 FALSE
alternative	양측검정, 단측검정

# 03 T-test를 사용하는 이유

## T-test의 종류

### 1) One-sample(단일표본 검정)

- **한** 모집단의 **평균**에 대한 검정
- $H_0 : \mu = \text{특정값}$

### 2) Paired(대응표본 검정) - [Before and After]

- **종속적인 짝**(쌍)을 이룬 두 변수간 **평균의 차이**에 대한 검정
- $H_0 : \mu_1 = \mu_2$

### 3) Two-sample(독립표본 검정)

- **독립적인 두** 모집단의 **평균의 차이**에 대한 검정
- $H_0 : \mu_1 = \mu_2$

# 03 T-test를 사용하는 이유

## T-test 정리



Q1. 약 복용 전후

Q2. 비타민 부원전체 중간고사 평균?

Q3. 비타민의 A,B 두조의 중간고사 평균 차이?

A1. – Paired

A2. – One sample

A3. – Two sample

## 최종 정리

표본집단 **1개**에 대한 추정이다 = “One sample T.” 사용

표본집단이 2개인데, **독립**적이다 = “Two sample T.” 사용

표본집단이 2개인데, **종속**적이다 = “Paired T.” 사용

# 03 T-test를 사용하는 이유

## T-test 변수 설명

**1)paired** = 대응 표본(한 집단)[=T] vs 독립 표본(두 집단)[=F]  
**Cf)** T-test의 3번 조건 – 집단이 2개인가? – 표본집단이 2개인가?  
->여기서의 집단은 "모"집단이 아닌, "**표본**"집단을 이야기한다.

**3)alternative** = Two sided vs One sided(less, greater)  
->two.sided = A B 집단이 서로 같은지  
->less = A집단이 B집단보다 작은지 ("**A**"<B)  
->greater = A집단이 B집단보다 큰지 ("**A**">B)

# 03 T-test를 사용하는 이유

## T-test 변수 설명

2) **var.equal** = 두 집단의 분산이 같은지[=T] vs 다른지[=F]

```
> #sample vector1  
> var_test_vector1<-c(75,67,78,81,53,71,71,55,40,78,76,42,67,98,59,63,84,50,67,80,83)  
> #sample vector2  
> var_test_vector2<-c(58,81,77,80,76,63,54,64,85,54,70,71,71,55,40,78,76,100,51,42,63,61,82,57,48)
```

```
> #분산이 같은지 확인  
> var.test(var_test_vector1,var_test_vector2)
```

F test to compare two variances

```
data: var_test_vector1 and var_test_vector2  
F = 1.0085, num df = 20, denom df = 24, p-value = 0.974  
alternative hypothesis: true ratio of variances is not equal to 1  
95 percent confidence interval:  
 0.433347 2.428065  
sample estimates:  
ratio of variances  
 1.008516
```

= P-value>0.05 = Ho 기각 X = 두 집단의 분산의 차이가 없다.

# 03 T-test를 사용하는 이유

## T-test 예시 with R

▶ “동아리원”들이 “비타민”에 들어온 후 “R성적 향상 여부” 확인

1) 모집단 = 비타민 동아리원들 (동일 집단 = “대응”표본)

2)  $H_0$  = 비타민에 입단 전 R 성적 == 비타민 입단 후 R 성적

3)  $H_a$  = 비타민에 입단 전 R 성적 != 비타민 입단 후 R 성적

```
> #비타민에 다니기 전의 학생 점수
> before_study <- c(34,76,76,63,73,75,67,78,81,53,58,81,77,80,43,65,76,63,54,64,85,54,70,71,71,
+                   55,40,78,76,100,51,93,64,42,63,61,82,67,98,59,63,84,50,67,80,83,66,86,57,48)
>
> #비타민에 다닌 후 학생 점수
> after_study <- c(74,87,89,98,65,82,70,70,70,84,56,76,72,69,73,61,83,82,89,75,48,72,80,66,82,
+                  71,49,54,70,65,74,63,65,101,82,75,62,83,90,76,87,90,78,63,59,79,74,65,77,74)
```

## 03

## T-Test 함수

## &lt;이분산&gt;

&lt; 이분산 가정 &gt;

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{V_x}{n_x} + \frac{V_y}{n_y}\right)}}$$

 $\sigma_1^2 \neq \sigma_2^2$  일 때  $\mu_1 - \mu_2$ 의 검정통계량

$$t = \frac{(E(\bar{X}_1) - E(\bar{X}_2)) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

## &lt;등분산&gt;

&lt; 등분산 가정 &gt;

$$t = \frac{\bar{x} - \bar{y}}{\sigma_p \times \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

 $S_p$ =합동표본분산(pooled sample variance) $\sigma_1^2 = \sigma_2^2$  일 경우

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

자유도  $v = n_1 + n_2 - 2$ 

F검정(등분산,이분산 검정) -&gt; T검정

## 03

## T-Test 함수

&lt;paired&gt; = 종속적

짝을 이룬 두 표본의 대응 비교 (paired comparison)

	1	2	...	n	평균 (mean)	표준편차 (sd)
Paired Sample 1	$X_1$	$X_2$	...	$X_n$		
Paired Sample 2	$Y_1$	$Y_2$	...	$Y_n$		
$D_i = X_i - Y_i$	$D_1$	$D_2$	...	$D_n$	$\bar{D}$	$S_D$

짝을 이룬 두 표본  $X, Y$  에서  $\bar{D} = \frac{\sum_{i=1}^n (X_i - Y_i)}{n}$  일 때,

t-검정 검정통계량  $T = \frac{\bar{D}}{S_D / \sqrt{n}}$

F검정(등분산,이분산 검정) -&gt; T검정



# 03 T-test를 사용하는 이유

## 양측검정("two-sided")

```
> #t-검정 수행(양측검정)  
> t.test(before_study, after_study, paired = TRUE)
```

Paired t-test

data: before\_study and after\_study

t = -2.1129, df = 49, p-value = 0.03973

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.6285348 -0.2914652

sample estimates:

mean of the differences

-5.96

2)  $H_0$  = 비타민에 입단 전 R 성적 == 비타민 입단 후 R 성적

3)  $H_a$  = 비타민에 입단 전 R 성적 != 비타민 입단 후 R 성적

= P-value < 0.05 =  $H_0$  기각 ○ = 비타민 입단 후 성적 차이가 있다.

# 03 T-test를 사용하는 이유

## 단측검정("one-sided" - "less")

```
> #t-검정 수행(단측검정-less)
> t.test(before_study, after_study, paired = TRUE, alternative = 'less')
```

Paired t-test

```
data: before_study and after_study
t = -2.1129, df = 49, p-value = 0.01986
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -1.230848
sample estimates:
mean of the differences
      -5.96
```

2)  $H_0$  = 비타민에 입단 전 R 성적 == 비타민 입단 후 R 성적  
3)  $H_a$  = 비타민에 입단 전 R 성적 < 비타민 입단 후 R 성적

=  $P\text{-value} < 0.05$  =  $H_0$  기각 ○ = 비타민 입단 후 성적 차이가 있다.

# 03 T-test를 사용하는 이유

## 단측검정("one-sided" - "greater")

```
> #t-검정 수행(단측검정-greater)
> t.test(before_study, after_study, paired = TRUE, alternative = 'greater')
```

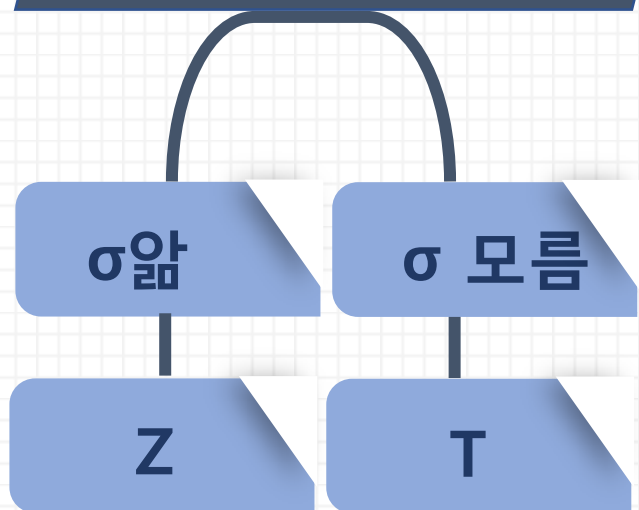
Paired t-test

```
data: before_study and after_study
t = -2.1129, df = 49, p-value = 0.9801
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -10.68915      Inf
sample estimates:
mean of the differences
      -5.96
```

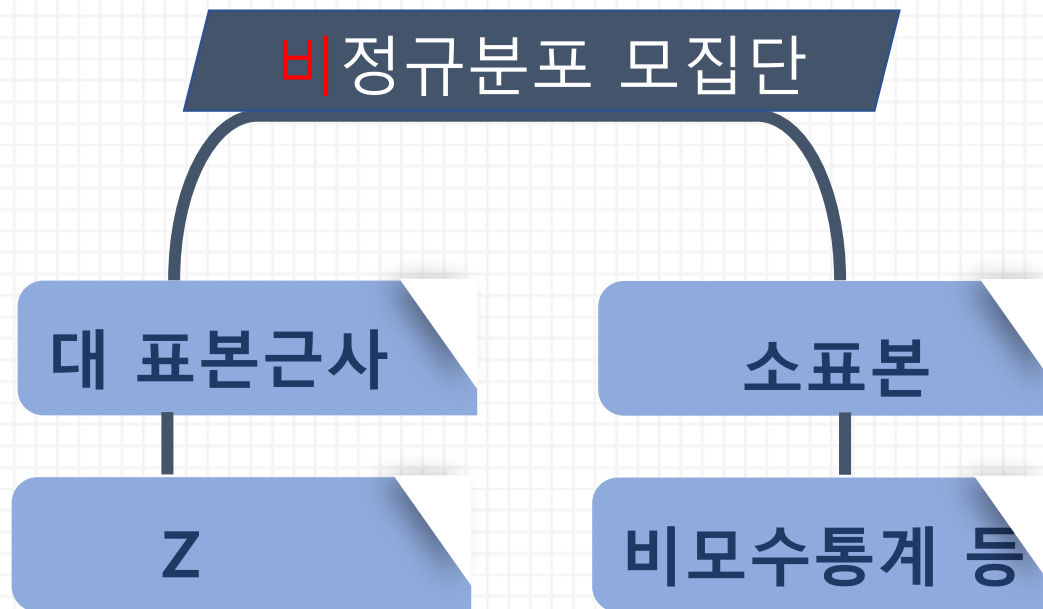
2)  $H_0$  = 비타민에 입단 전 R 성적 == 비타민 입단 후 R 성적  
3)  $H_a$  = 비타민에 입단 전 R 성적 > 비타민 입단 후 R 성적

=  $P\text{-value} < 0.05$  =  $H_0$  기각 ○ = 비타민 입단 후 성적 차이가 있다.

## 정규분포 모집단



## 비정규분포 모집단



1) 정규분포를 따르는가? -> 2)  $\sigma$ 를 아는가? -> 3) 분산이 같은가?(F분포)