

# 통계 이론 정리

2017272043 이성진

# 변수의 분류

## 원인에 해당하는 변수

독립변수 Independent variable

설명변수 Explanatory variable

예측변수 Predictor variable

위험인자 Risk factor

공변량 Covariate

연속형 자료

요인 Factor

범주형 자료



## 결과에 해당하는 변수

종속변수 Dependent variable

반응변수 Response variable

결과변수 Outcome variable

표적변수 Target variable

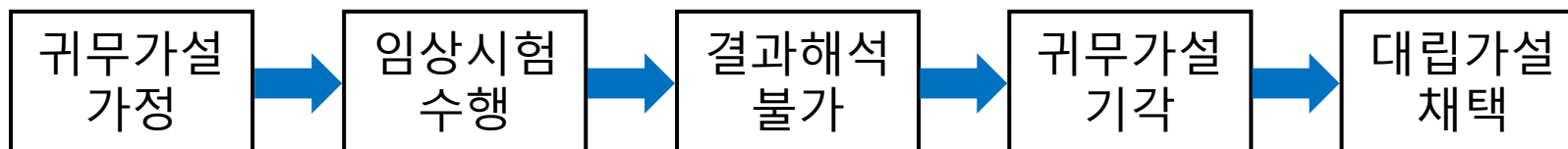
# 자료의 분류

범주형 자료 Categorical data	명목 척도 Nominal scale	범주			
	순위 척도 Ordinal scale	범주	순위		
연속형 자료 Numerical data	간격 척도 Interval scale	범주	순위	같은 간격	
	비 척도 Ratio scale	범주	순위	같은 간격	절대 영점

# 가설 검정 방법

"다른 상황을 생각하게 하는 현저한 근거가 없는 한 현상적인 모든 차이는 0(무)이다."

귀무가설 $H_0$	효과(혹은 차이)가 없다.
대립가설 $H_1$	효과(혹은 차이)가 있다.



$P\text{-value} < 0.05$



$p\text{-value} < 0.05$  : '귀무가설 가정 시 이 현상이 관찰될 확률이 5% 미만' 이라는 뜻

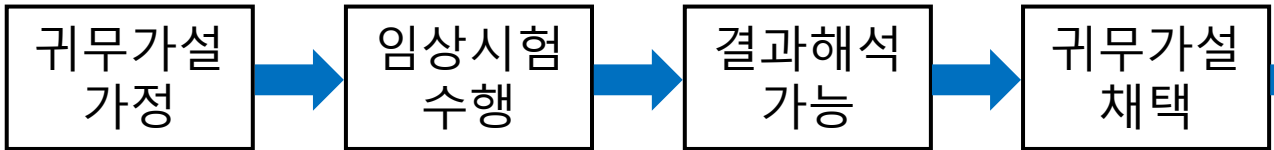
# 동등성 검정

Ex) 치료법 A와 B를 적용한 환자의 결과 비교

	반응	반응 없음	전체
치료군 A	45	5	50
치료군 B	40	10	50



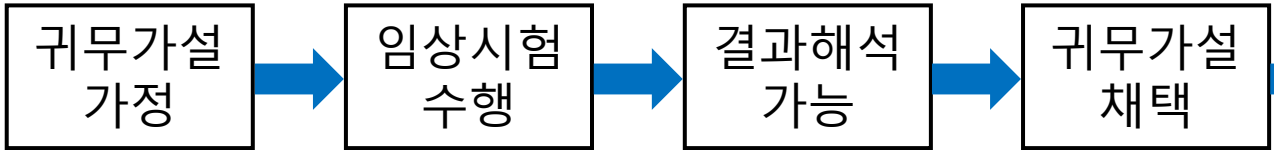
교차분석 수행  
*p-value = 0.161*



"두 치료법 A와 B는 효과의 차이가 없다?"

*P-value = 0.161*

- ※ 귀무가설을 기각하지 못했다고 하여 귀무가설이 항상 옳은 것은 아님
- ※ 동등성 검정 시 증명하고자 하는 명제인 '차이가 없다'가 대립가설이 된다




"두 치료법 A와 B는 효과의 차이가 있다!"

*P-value = 0.161*

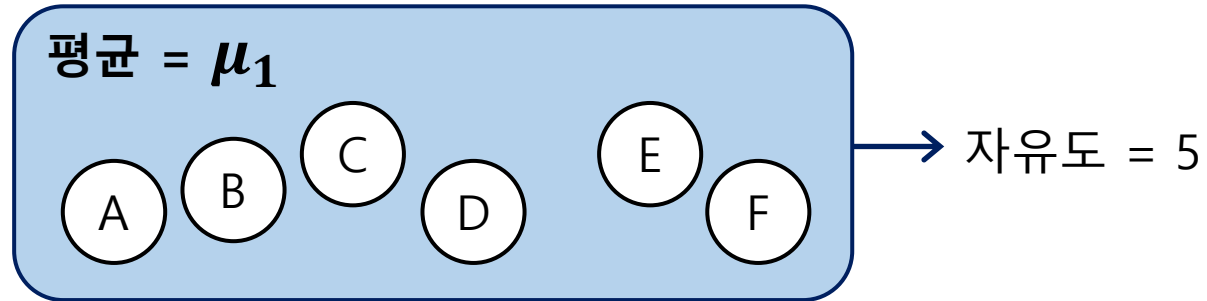
# 가설 검정 시 오류

		실제 진리	
		실제로 효과 없음	실제로 효과 있음
검정 결과	실험결과 효과 없음 귀무가설 채택	참	오류 제2종 오류( $\beta$ )
	실험결과 효과 있음 귀무가설 기각	오류 제1종 오류( $\alpha$ )	참 검정력( $1-\beta$ )
		$p\text{-value} < 0.05$	



# 자유도

'실질적으로 독립인 값들의 개수'



	당뇨	정상	전체
고혈압	<i>a</i>	<i>b</i>	20
정상	<i>c</i>	<i>d</i>	80
전체	25	75	100

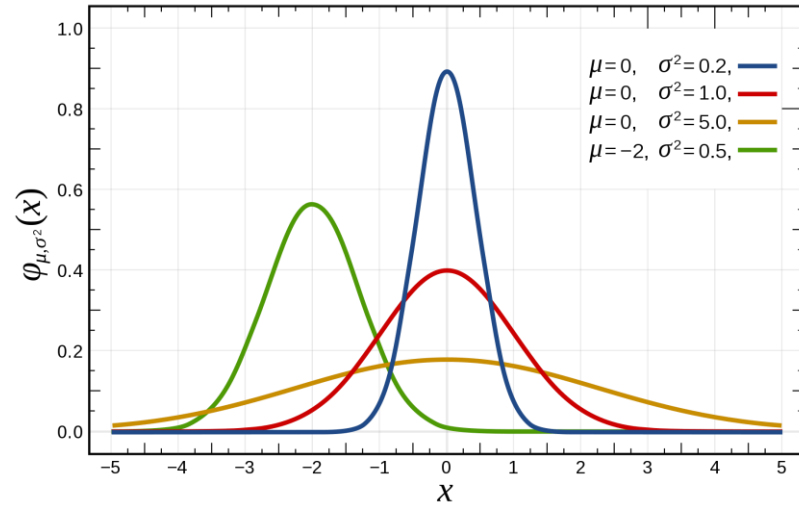
→ 자유도 = 1

	당뇨	내당능장애	정상	전체
고혈압	<i>a</i>	<i>b</i>	<i>c</i>	20
정상	<i>d</i>	<i>e</i>	<i>f</i>	80
전체	25	25	50	100

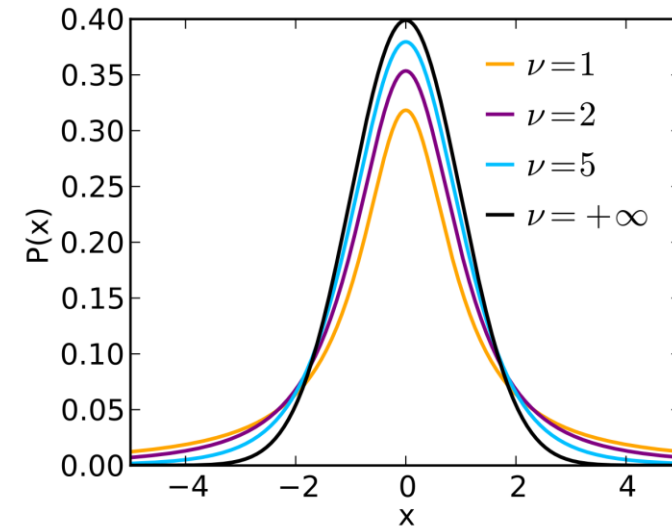
→ 자유도 = 2

# 분포

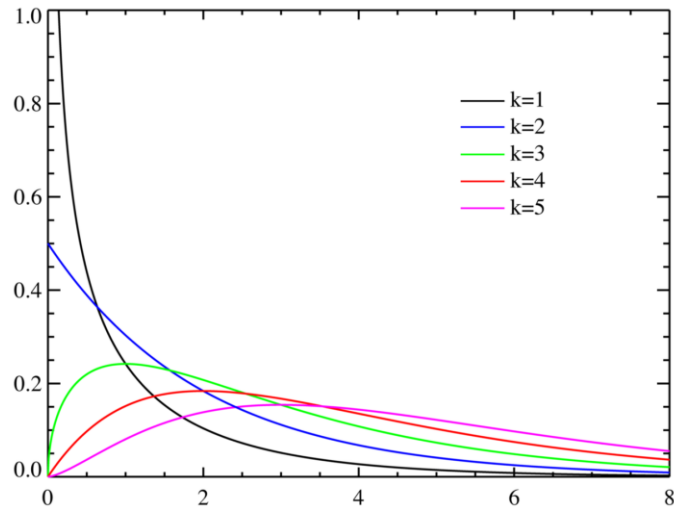
## 정규분포 $N(\mu, \sigma^2)$



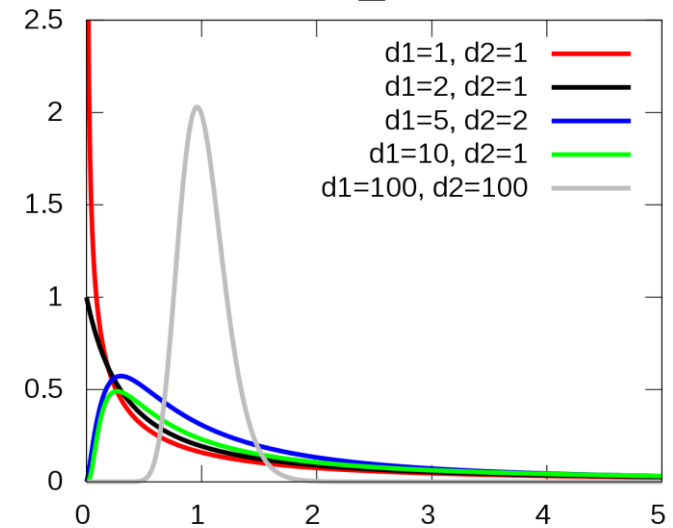
## t-분포



## 카이제곱 분포



## F-분포





# 분포와 검정통계량

- 통계적 가설검정에는 특정 확률 분포 이용 (ex) 독립표본 T검정 → t-분포, ANOVA → F분포)
- 가설검정 시 검정통계량(test statistic) 계산
- 가정된 분포에서 관찰된 표본(검정통계량)이 발견될 확률 = 곡선 아래 면적

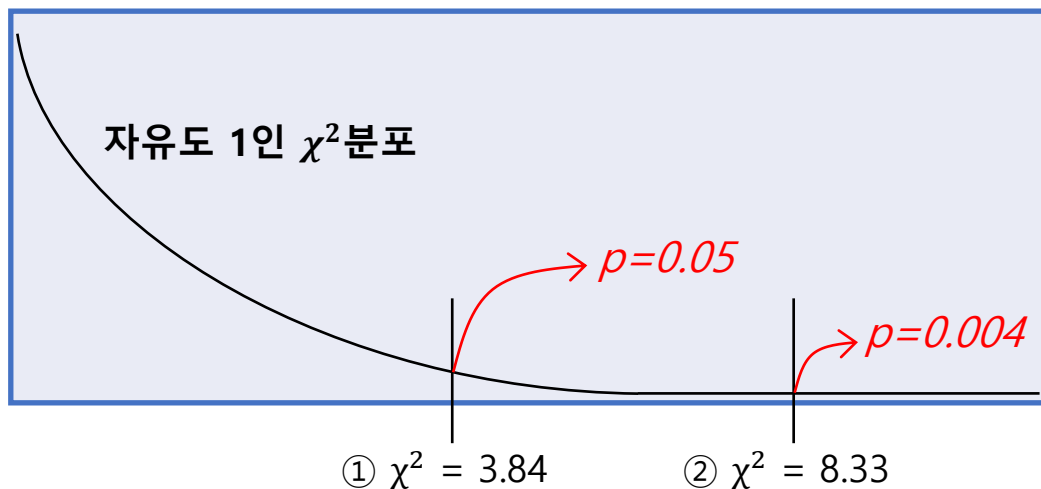
Ex. 당뇨와 비만 사이 연관성

	당뇨	정상	전체
비만	10	10	20
정상 체중	15	65	80
전체	25	75	100

2×2 분할표 →  $\chi^2$  검정 통계량은 자유도 1인  $\chi^2$  분포

$$\text{검정통계량 } \chi^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = 8.33$$

$\chi^2 = 8.33 \rightarrow p\text{-value} = 0.004 \rightarrow$  귀무가설 기각 대립가설 채택



귀무가설  $H_0$

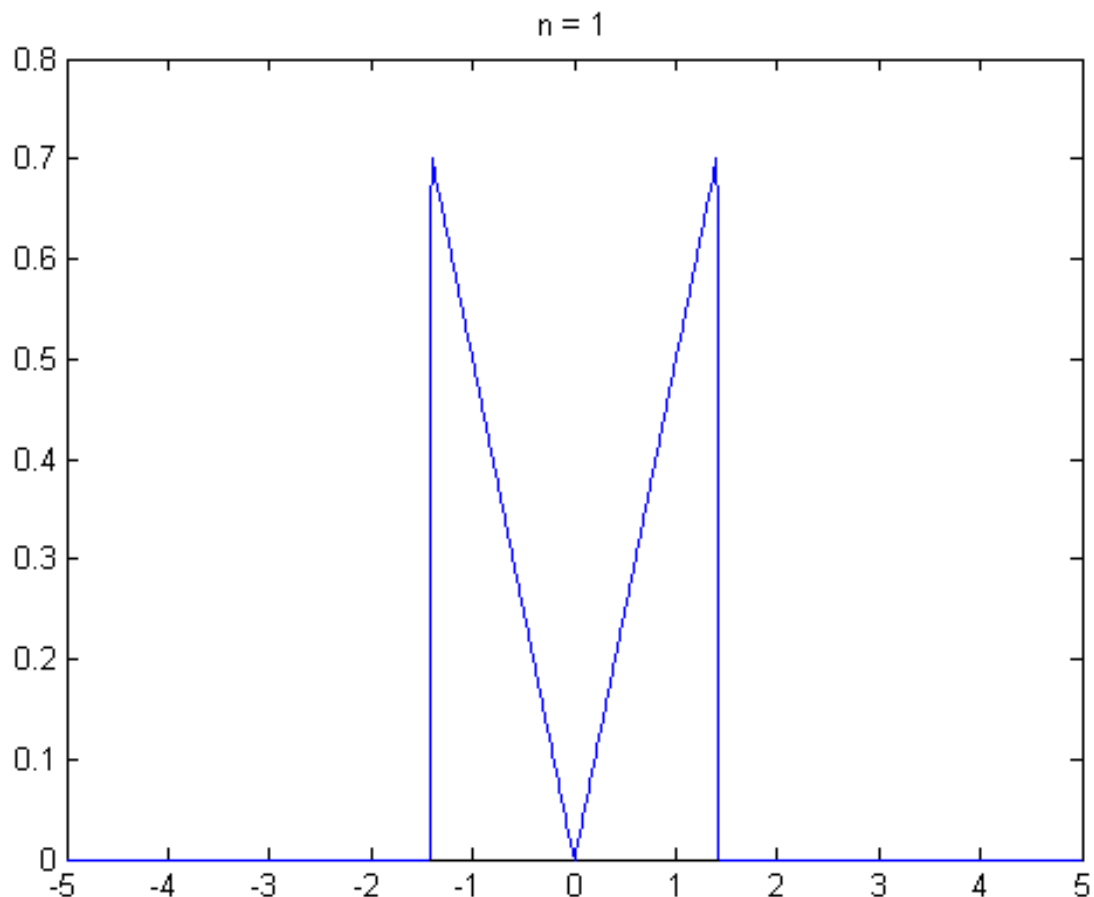
당뇨와 비만 사이에 연관성이 없다.

대립가설  $H_1$

당뇨와 비만 사이에 연관성이 있다.

# 중심극한정리(Central limit theorem)

'모집단으로부터 무작위로 표본을 추출할 때 표본의 크기가 충분히 크다면  
표본의 합 또는 평균의 히스토그램은 정규분포 곡선에 수렴한다'



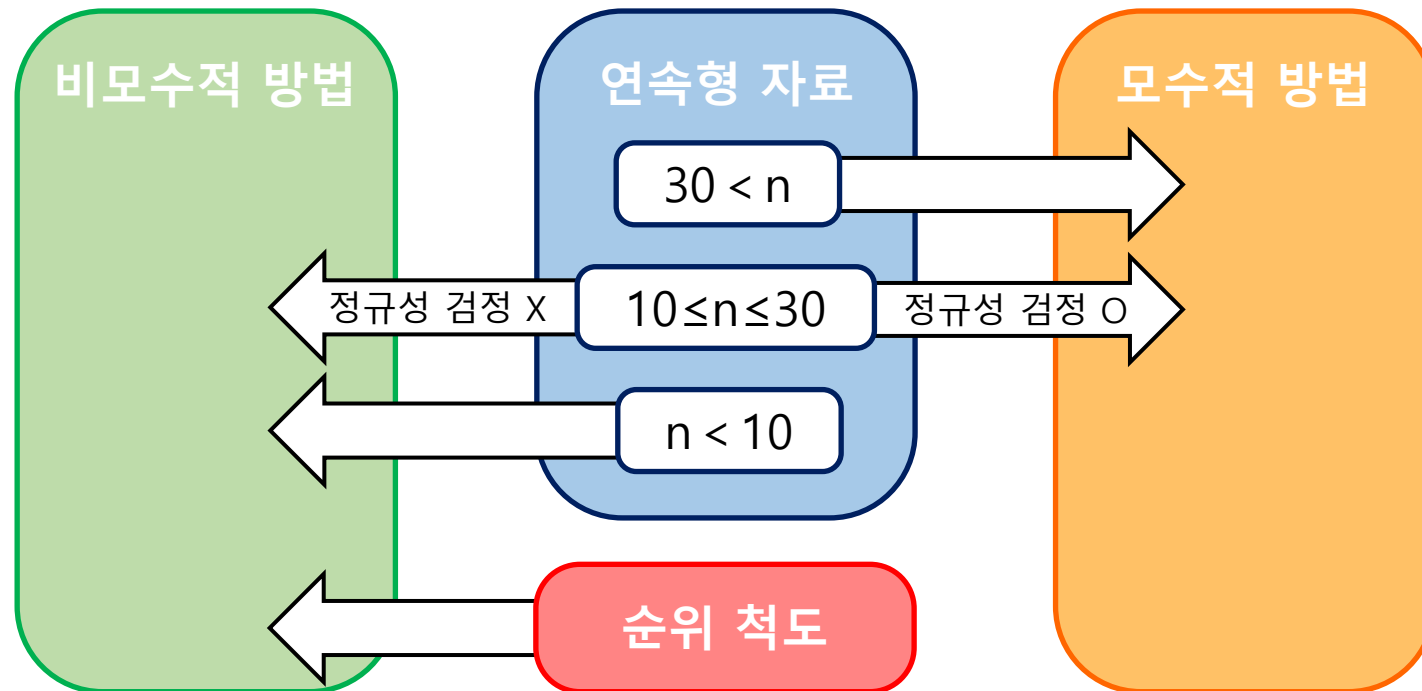
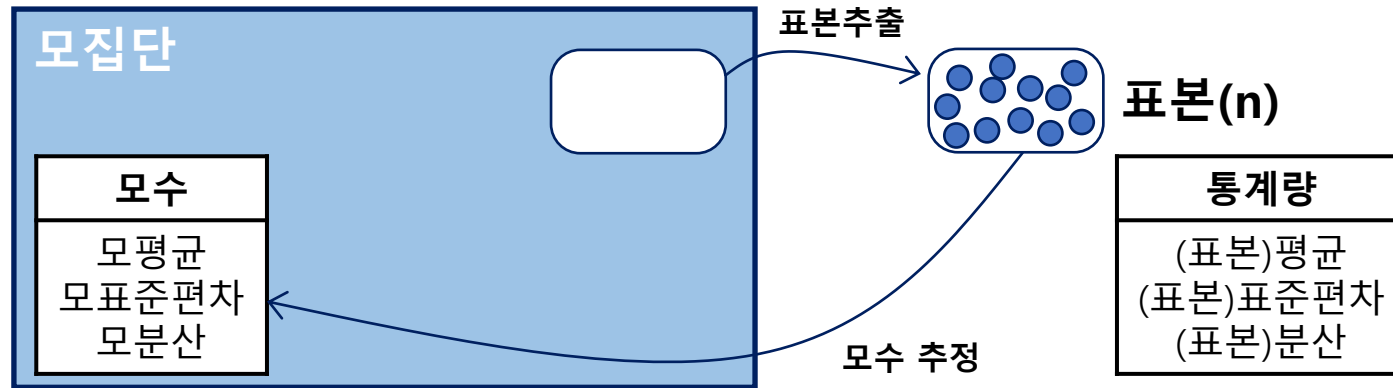
## <표본 크기와 정규 분포>

표본 평균이 정규 분포를 따른다고 가정이 가능한 최소 표본 크기	$n=30$
정규성 검정을 통해 표본 평균이 정규 분포를 따르는지 확인할 표본 크기	$10 \leq n \leq 30$
표본 평균이 정규 분포를 따른다고 가정할 수 없는 표본 크기	$n < 10$

## <정규성 검정의 가설 설정>

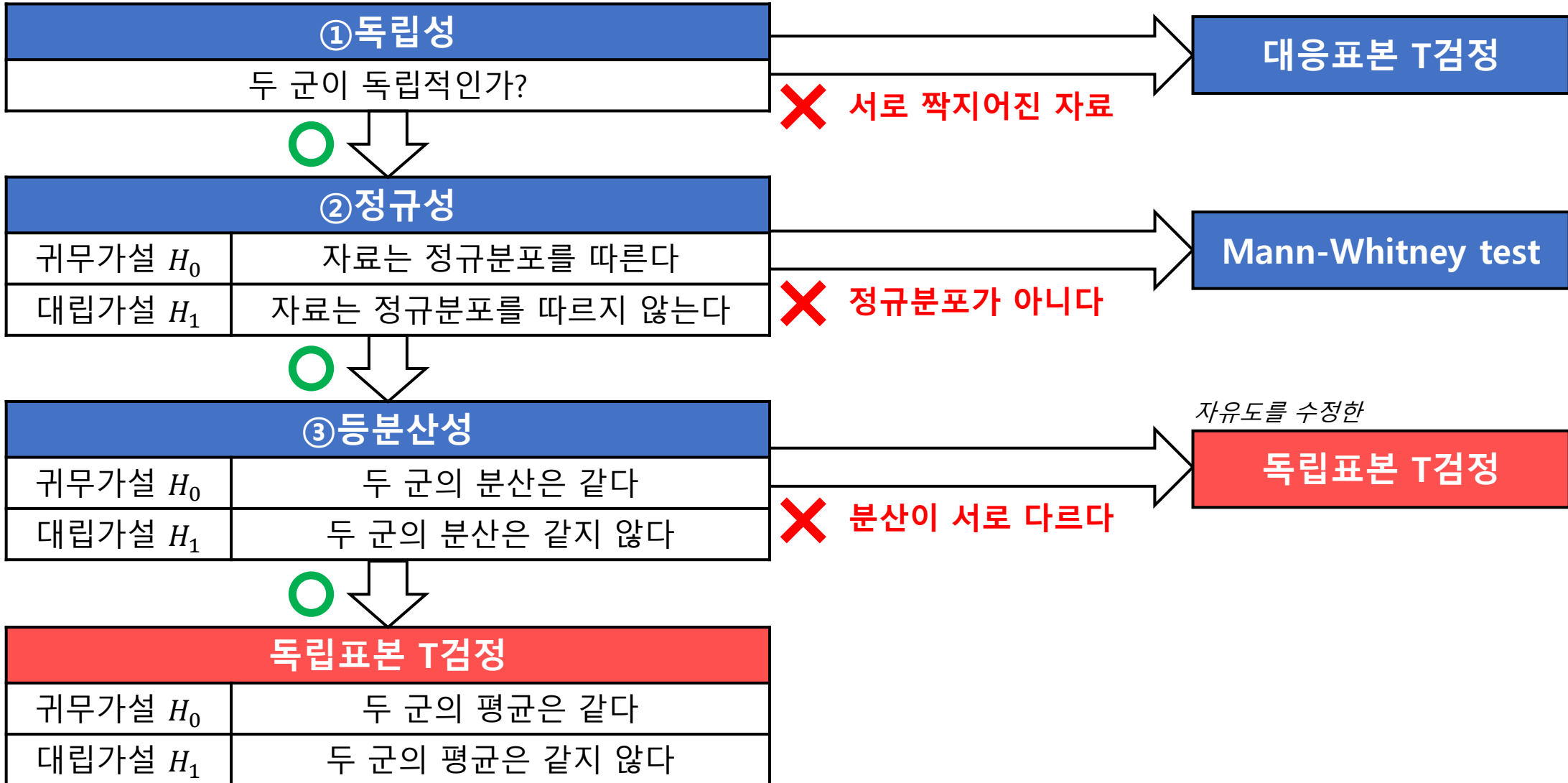
귀무가설 $H_0$	자료는 정규분포를 따른다	$p \geq 0.05$
대립가설 $H_1$	자료는 정규분포를 따르지 않는다	

# 모수적 방법 & 비모수적 방법



# 독립표본 T검정

결과변수가 연속형인 독립된 두 군의 크기를 비교하는 모수적 방법



## 다중 비교의 문제(Multiple Comparison problem)

A군 VS B군  $p > 0.05$

B군 VS C군  $p > 0.05$

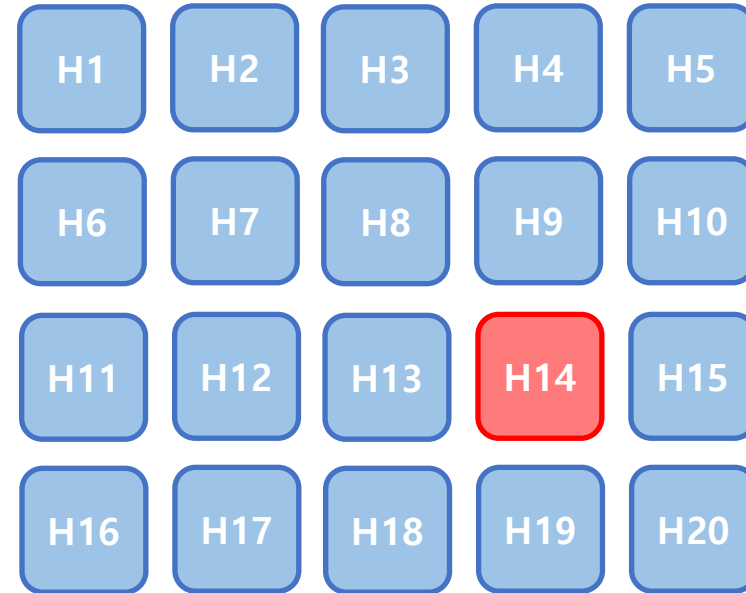
C군 VS A군  $p < 0.05$



'통계적으로 유의한 차이가 있다?'

전체 유의수준 :

$$1 - (1 - 0.05)^3 \approx 0.143(14.3\%)$$



전체 유의수준 :

=  $1 - (\text{모두 효과가 있을 확률})$

$$= 1 - (1 - 0.05)^{20}$$

$$\approx 0.642 (64.2\%) \gg 0.05$$

= 적어도 하나는 효과가 있다고 결론 낼 확률 (제1종 오류)

# 독립된 세 군 이상의 크기 비교 방법

## STEP1 일원배치 분산분석

$$\text{A군} = \text{B군} = \text{C군}$$

귀무가설  $H_0$

세 군의 크기는 모두 같다

$$\text{A군} \neq \text{B군} \text{ OR } \text{B군} \neq \text{C군} \text{ OR } \text{C군} \neq \text{A군}$$

대립가설  $H_1$

적어도 한 쌍은 크기가 다르다

## STEP2 사후분석

$$\text{A군} = \text{B군}$$

$$\text{B군} = \text{C군}$$

$$\text{C군} = \text{A군}$$

귀무가설  $H_0$

두 군의 평균은 같다

$$\text{A군} \neq \text{B군}$$

$$\text{B군} \neq \text{C군}$$

$$\text{C군} \neq \text{A군}$$

대립가설  $H_1$

두 군의 평균은 같지 않다

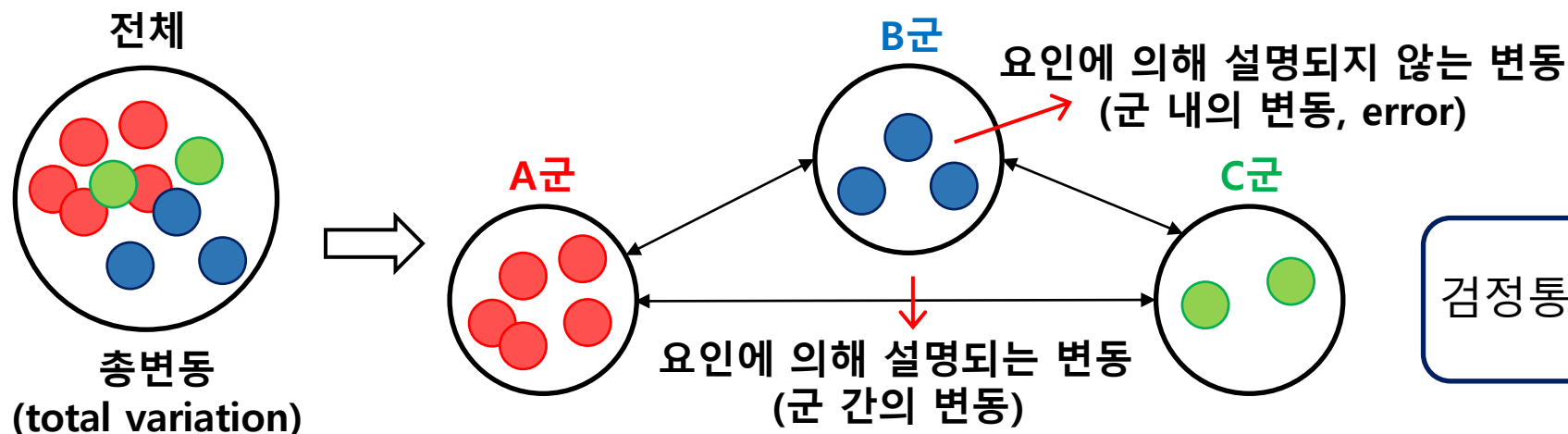
### 유의수준 보정

ex) Bonferroni's method (유의수준 =  $\frac{5\%}{\text{검정의 횟수}}$ )

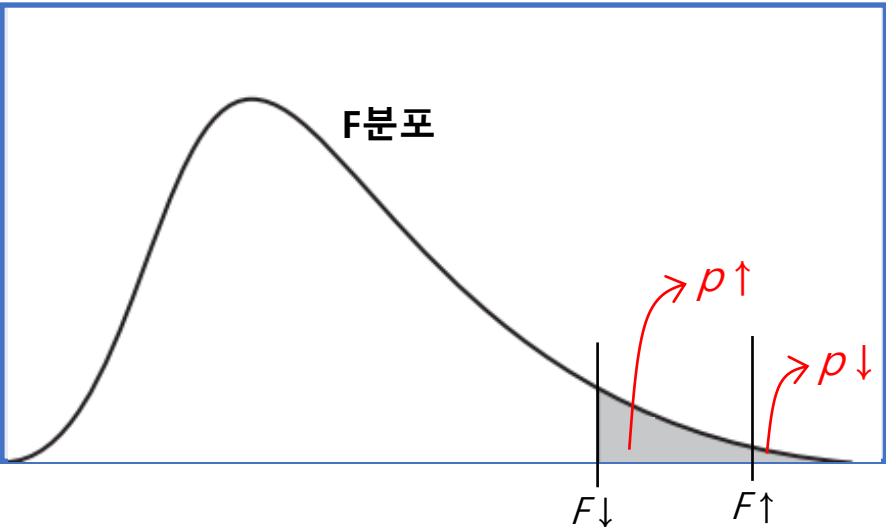
# 일원배치 분산분석(ANOVA)

변동(variation) : 자료들 간의 크기의 차이 혹은 변화량의 집체적인 표현

요인(factor) : 전체 자료들을 구분짓는 요소 (군)



$$\text{검정통계량 } F = \frac{\text{군 간의 변동의 평균}}{\text{군 내의 변동의 평균}}$$



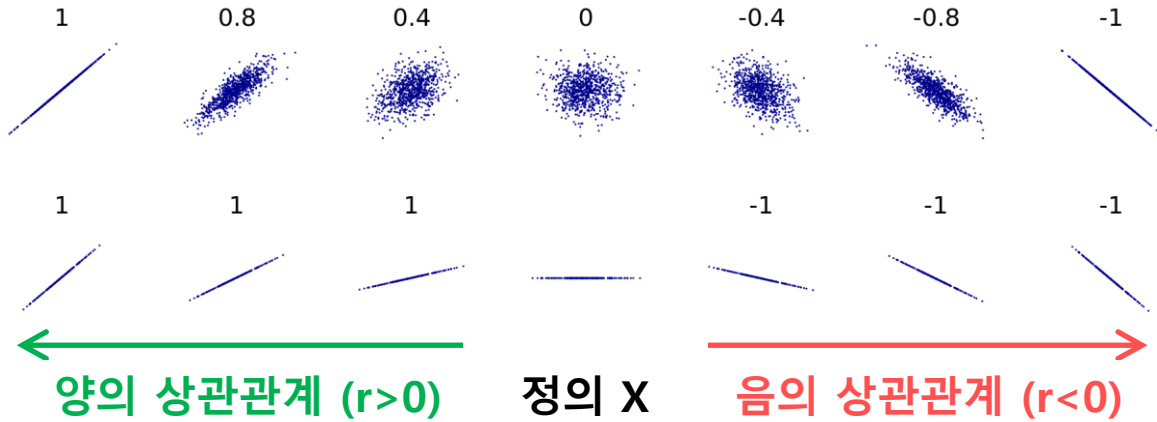
분산분석표

요인	제곱합(변동)	자유도	제곱합의 평균	F 통계량	유의확률
군	SSG(군 간의 변동)	$g-1$	$MSG=SSG/(g-1)$	$F=MSG/MSE$	$p\text{-value}$
오차	SSE(군 내의 변동)	$n-g$	$MSE=SSE/(n-g)$		
전체	TSS(총변동)	$n-1$			

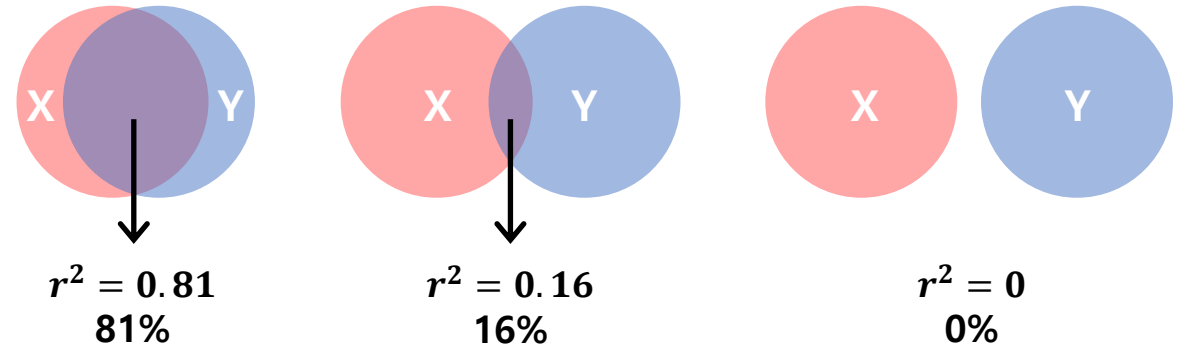
g(군의 수)  
n(총개체 수)

# Pearson의 상관분석

'두 연속형 변수의 상관 정도에 대해 알려주는 분석법'



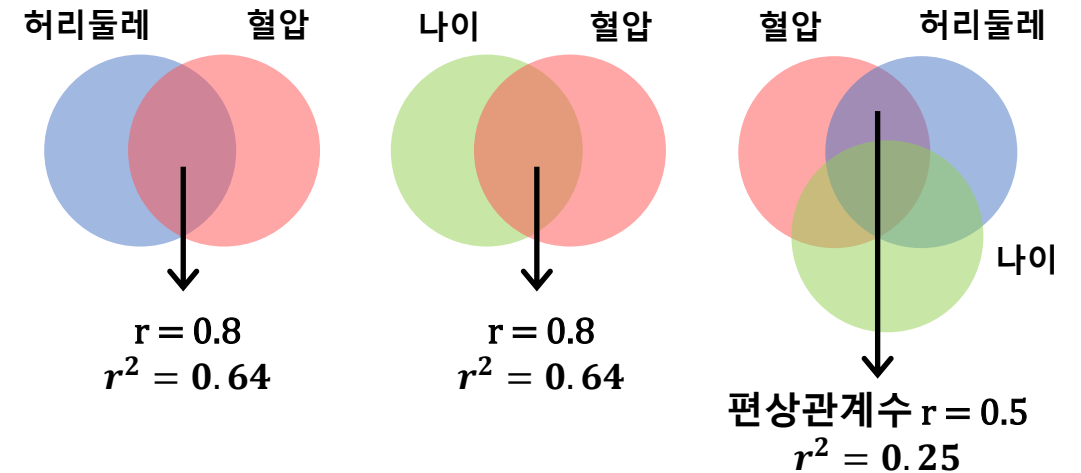
설명력  $r^2$  : 두 변수 사이의 선형 관계의 정도 설명



## 상관분석의 가설 설정

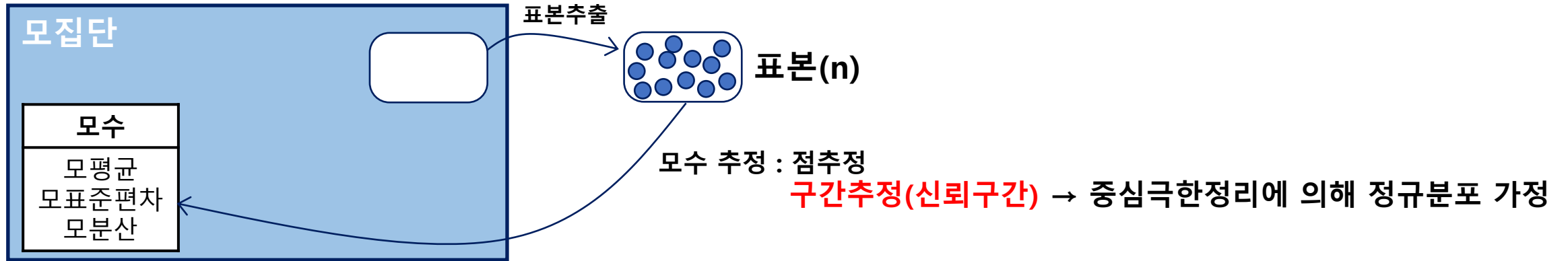
귀무가설 $H_0$	두 변수는 선형의 상관관계가 없다 ( $r=0$ )
대립가설 $H_1$	두 변수는 선형의 상관관계가 있다 ( $r \neq 0$ )

## 편상관분석(partial correlation)





# 95% 신뢰구간



$$\text{모평균의 95\% 신뢰구간} = \bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}}$$

(표본평균  $\bar{X}$ , 표본 표준편차  $s$ , 표본의 크기  $n$ )

$$\text{모비율의 95\% 신뢰구간} = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

(표본의 관심사건의 비율  $p$ , 표본의 크기  $n$ )

## 독립 표본 T 검정에서의 95% 신뢰구간

