

# 기초통계 과정

## - 엑셀 활용 -

---

스마트미디어인재개발원

---

[smhrd@smhrd.or.kr](mailto:smhrd@smhrd.or.kr)

---

062.655.3507

---

# 목차

---

1. 기초 통계 이론

---

2. 기초 통계 분석

---

# 목차

---

## 1. 기초 통계 이론

1.1 통계학 및 데이터

1.2 정규분포

1.3 데이터 요약

1.4 검정 통계

---

## 2. 기초 통계 분석

---

# [기초 통계 이론] 들어가기에 앞서...

**학습목표** ✓ 데이터에 대한 이론적 이해와 통계 분석을 하는데 필요한  
검정 통계 및 가설에 대한 이론을 이해하고 숙지한다.

**학습내용** ✓ 통계 이론을 통해 데이터에 대해서 알아보자  
✓ 검정 통계란 무엇이며, 필요한 통계량은 무엇인가?  
✓ 가설이란 무엇이며, 어떤 방식으로 가설을 검정하는가?

# 1.1 통계학 및 데이터

**정의** ✓ 통계학은 데이터를 통하여 분석하고 분석결과를 통해 정보를 제공하는 분야

**목적** ✓ 통계학을 시작함에 앞서 통계적 목적에 따른 통계학의 구분과 분석에 사용되는 데이터의 구성 및 형태를 파악하기 위함

**Focus!**

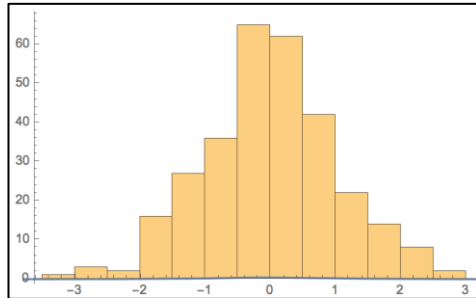
1. 통계학은 통계적 목적에 따라 어떻게 구분되는가?
2. 데이터의 구성은 어떻게 정의하는가?
3. 데이터의 형태는 어떻게 구분되는가?

# Focus 1. 통계학은 통계적 목적에 따라 어떻게 구분되는가?

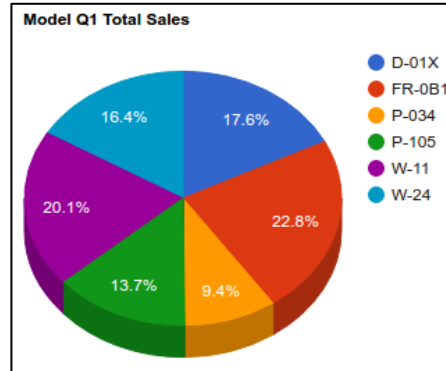


# 기술 통계학

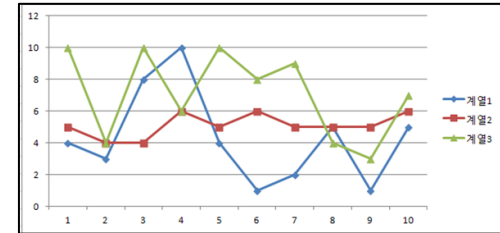
## ■ 히스토그램



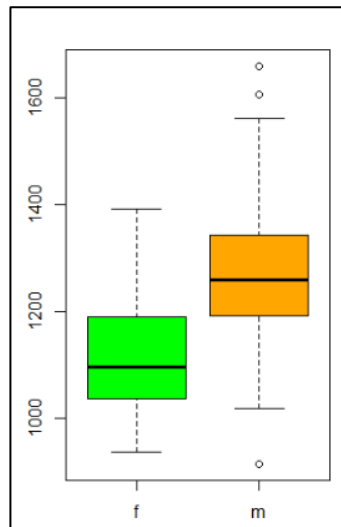
## ■ 원 그래프



## ■ 꺾은선 그래프



## ■ 상자 그래프



## ■ 평균, 사분위수 등

평균	3,0133	,06828
평균의 95% 신뢰구간	2,8757	
하한	3,1509	
상한		
5% 절삭평균	3,0086	
중위수	3,0000	
분산	,210	
표준편차	,45806	
최소값	2,20	
최대값	3,80	
범위	1,60	
사분위수 범위	,80	
왜도	,069	,354
첨도	-1,137	,695

## ■ 줄기-잎 그래프

줄기	잎			
6	5	8	2	
7	1	4	1	0 9
8	2	7	2	5
9	8	2	3	1

# 추측 통계학

## ■ 선거 출구 조사

전체 유권자들(모집단)이 후보자들을 지지하는 지지율(모수)를 알기 위해서, 특정 유권자(표본)들을 추출하여 후보자별 지지율(통계량)이 얼마나 되는지를 조사하는 것

- ✓ 모집단 : 정보를 알고자 하는 대상들의 전체 집합
- ✓ 표본 : 모집단에 관한 정보를 얻기 위해서 사용되는 모집단의 일부분
- ✓ 모수 : 모집단의 특성을 나타내는 값
- ✓ 통계량 : 표본들의 특성을 나타내는 값



## Focus 2. 데이터의 구성은 어떻게 정의하는가?

개체

변수

PassengerId	Survived	Pclass	Sex
1	0	3	male
2	1	1	female
3	1	3	female
4	1	1	female
5	0	3	male
6	0	3	male
7	0	1	male
8	0	3	male
9	1	3	female

관측값

### ■ 변수

데이터를 저장한 이름으로 데이터의 속성을 표현하며 PassengerId(데이터의 넘버링), Survived(생존), Pclass(클래스), Sex(성별)와 같은 것을 의미함

- 설명(원인)변수, 반응(결과)변수 등이 있음

## Focus 2. 데이터의 구성은 어떻게 정의하는가?

개체

변수

PassengerId	Survived	Pclass	Sex
1	0	3	male
2	1	1	female
3	1	3	female
4	1	1	female
5	0	3	male
6	0	3	male
7	0	1	male
8	0	3	male
9	1	3	female

관측값

### ■ 개체

서로 다른 데이터를 구분해주는 값으로 중복되지 않는 값이며, PassengerId 변수에 입력된 데이터의 넘버링 값과 같은 고유한 값을 의미함

## Focus 2. 데이터의 구성은 어떻게 정의하는가?

개체

변수

PassengerId	Survived	Pclass	Sex
1	0	3	male
2	1	1	female
3	1	3	female
4	1	1	female
5	0	3	male
6	0	3	male
7	0	1	male
8	0	3	male
9	1	3	female

관측값

### ■ 관측값

변수에 대응되는 값으로 Sex변수에 대한 male, female, Pclass변수에 대한 1,2,3 등과 같은 값을 의미함  
(개체값도 관측값이 될 수 있음)

## Focus 3. 데이터의 형태는 어떻게 구분되는가?



## Focus 3. 데이터의 형태는 어떻게 구분되는가?

- ✓ 양적 데이터를 가지고 표로 정리할 경우 **전달력이 떨어질 수 있다?**

〈 A반 학생 키 데이터 〉

160	162	170	175
180	178	161	162
179	176	165	181
166	168	175	178
168	164	172	164

〈 표로 정리한 결과 〉

키 구간	인원 수
160~164cm	6명
165~169cm	4명
170~174cm	2명
175~179cm	6명
180cm 이상	2명
전체	20명

〈 숫자로 요약한 결과 〉

최소 키	160.0cm
평균 키	170.2cm
최대 키	181.0cm

- ✓ 질적 데이터를 가지고 숫자로 요약할 경우 **잘못된 결과가 전달 될 수 있다?**

〈 A반 혈액형 데이터 〉

1	4	3	2	1	1 : A형
2	3	3	1	4	2 : B형
4	1	2	1	3	3 : AB형
1	2	3	2	4	4 : O형

〈 숫자로 요약한 결과 〉

최소값	1 (A형)
평균값	2.4 (B형?)
최대값	4 (O형)

〈 표로 정리한 결과 〉

혈액형	인원 수
A형	6명
B형	5명
AB형	5명
O형	4명
전체	20명

## 1.2 정규분포

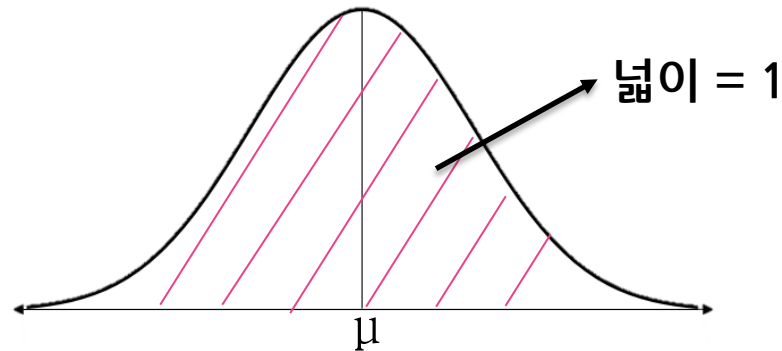
**정의** ✓ 데이터의 분포 형태가 **평균을 기준으로 대칭**에 가깝게 따를 때  
일반적으로 정규분포라 칭함  
(중심극한정리에 의하여 데이터가 충분히 클 경우 데이터의 분포는 정규분포를 따름)

**목적** ✓ 데이터를 분석함에 앞서 데이터 분포의 특성을 정확하게 이해하기 위해 정규분포의 특징 및 사례를 알아보고 경우에 따라 필요한 표준화된 정규분포까지 파악하기 위함

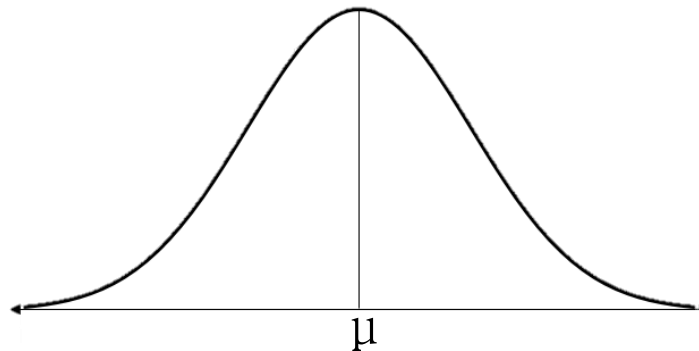
- Focus!**
1. 정규분포의 **특징**은 무엇인가?
  2. 정규분포를 활용한 **사례**가 있는가?
  3. **정규분포와 표준정규분포 차이점**은 무엇인가?
  4. **중심극한정리**란 무엇인가?

## Focus 1. 정규분포의 특징은 무엇인가?

- 정규분포 아래 부분의 면적 넓이는 1임

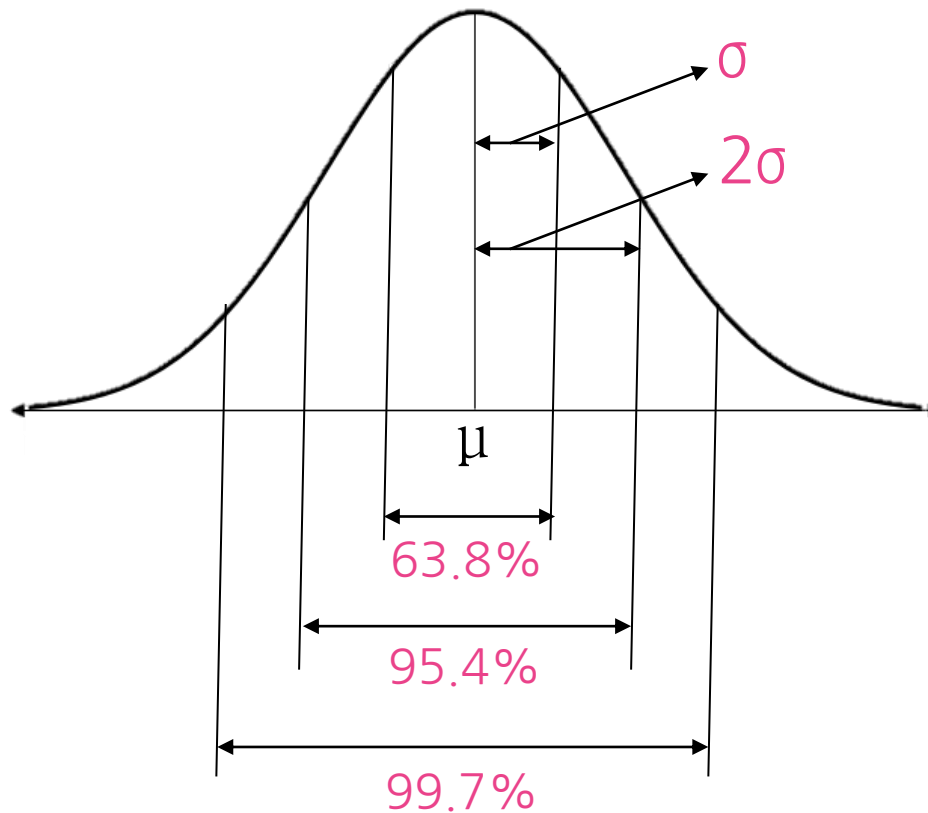


- 평균에 대하여 좌우대칭인 종모양



## Focus 1. 정규분포의 특징은 무엇인가?

- 평균 주변에 많이 몰려있으며 양 끝으로 갈수록 줄어들며, 표준편차( $\sigma$ )로 산포도 모양이 결정

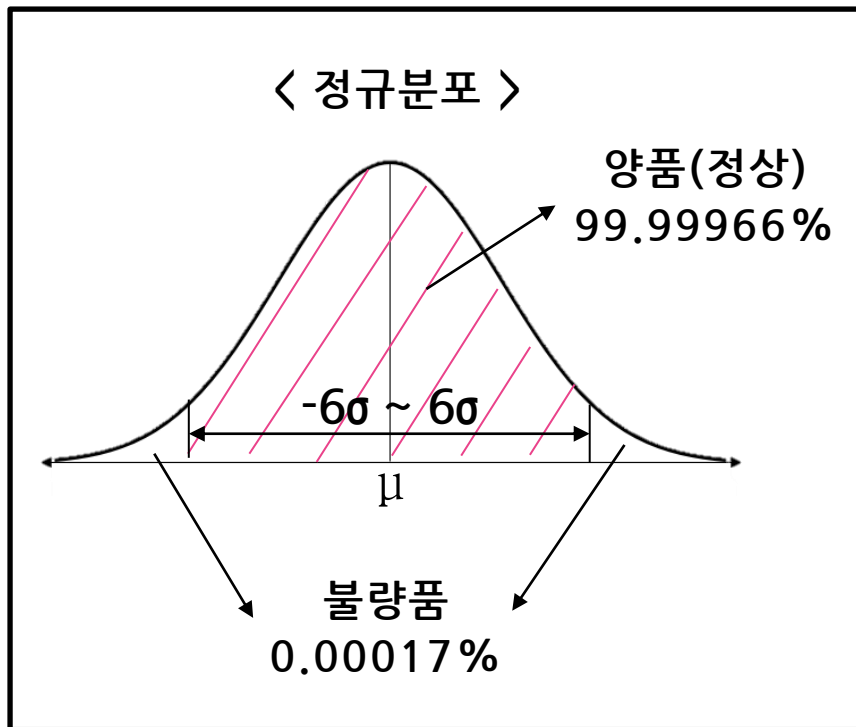




## Focus 2. 정규분포를 활용한 사례가 있는가?

### ■ 6시그마( $6\sigma$ )

기업에서 전략적으로 완벽에 가까운 제품이나 서비스를 개발하고 제 공하려는 목적으로 생성된 품질경영 기법



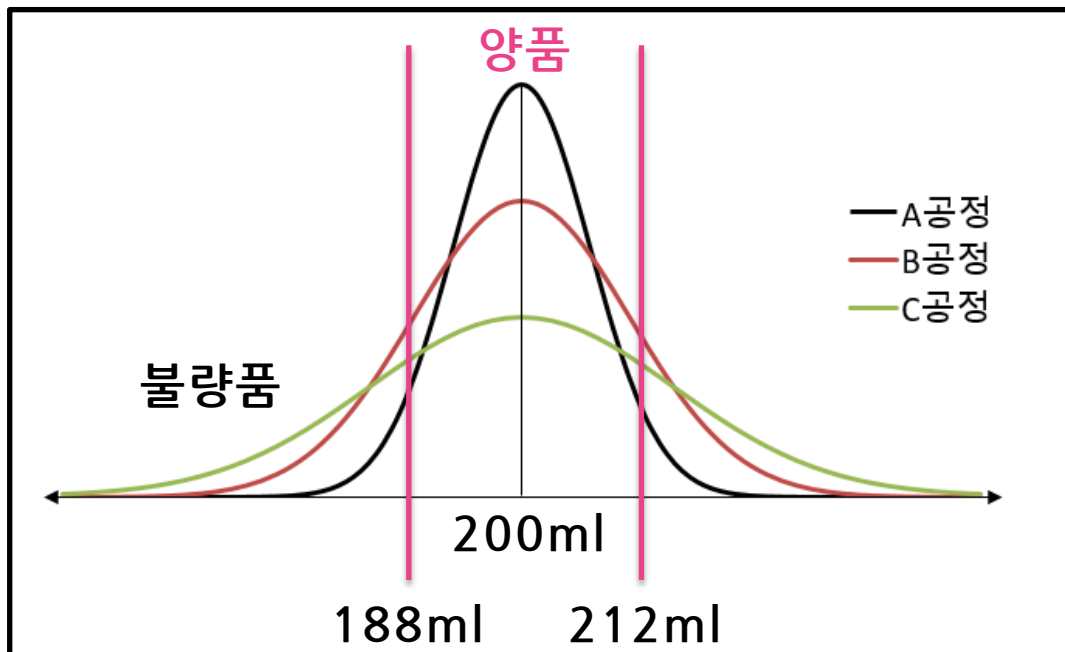
### ■ 해석

정규분포에서 평균( $\mu$ )을 중심으로 표준편차( $\sigma$ )의 6배 이내에 해당하는 구간만큼 양품(정상)을 생산할 수 있는 공정의 능력을 수치화한 것으로 100만개 중 3.4개의 불량품을 의미

## Focus 2. 정규분포를 활용한 사례가 있는가?

### ■ 6시그마( $6\sigma$ ) -> 예제

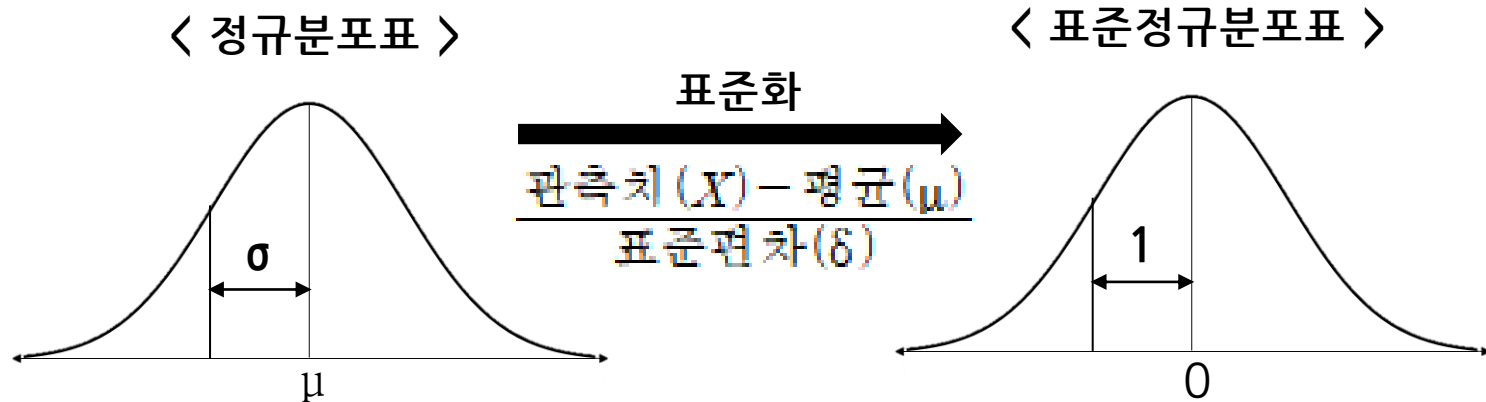
- 음료수를 병에 주입하는 3개의 공정 A,B,C에서 한 병에 들어가는 양의 분포 형태를 조사함
- 한 병에 200ml를 담는 것이 목표지만 주입량의 12ml의 편차를 허용함
- A공정의 6시그마 수준, B공정은 4시그마 수준, C공정은 3시그마 수준 사용



### ■ 각 공정별 표준편차( $\sigma$ )는?

- ✓ A공정 :  $6\sigma = 12 \rightarrow \sigma = 2$
- ✓ B공정 :  $4\sigma = 12 \rightarrow \sigma = 3$
- ✓ C공정 :  $3\sigma = 12 \rightarrow \sigma = 4$

## Focus 3. 정규분포와 표준정규분포 차이점은 무엇인가?



### ■ 표준화

평균을 0으로, 표준편차를 1로 만드는 것으로 비교하고자 하는 대상이 다른 산포도를 가지거나 다른 단위를 가질 때 상대적인 크기로 비교하기 위해 사용

- ✓ A학교에서 90점 맞은 학생과 B학교에서 130점 맞은 학생 중 누가 더 높은 점수인가?
- ✓ 키 5cm 차이와 몸무게 5kg 차이 중 어느 것이 더 큰 차이인가?

## Focus 3. 정규분포와 표준정규분포 차이점은 무엇인가?

- A학교에서 90점 맞은 학생과 B학교에서 130점 맞은 학생 중 누가 더 높은 점수인가?

A학교에서는 채점은 0~100점으로 하고, B학교에서는 0~200점으로 할 경우, A학교에서 90점 맞은 학생과 B학교에서 130점을 맞은 두 학생 중 어느 학생이 더 잘한 것인가?

(단, A학교의 평균은 70점, 표준편차는 5점이고 B학교의 평균은 100점 표준편차는 10점)

- A학교의 표준 점수?  $z = \frac{x - \mu}{\delta} = \frac{90 - 70}{5} = 4$

- B학교의 표준 점수?  $z = \frac{x - \mu}{\delta} = \frac{130 - 100}{5} = 3$

- A학교 학생이 B학교 학생보다 표준점수가 크므로 **A학교 학생이 B학교 학생보다 더 잘했다고 볼 수 있음**

## Focus 3. 정규분포와 표준정규분포 차이점은 무엇인가?

- 키 5cm 차이와 몸무게 5kg 차이 중 어느 것이 더 큰 차이인가?

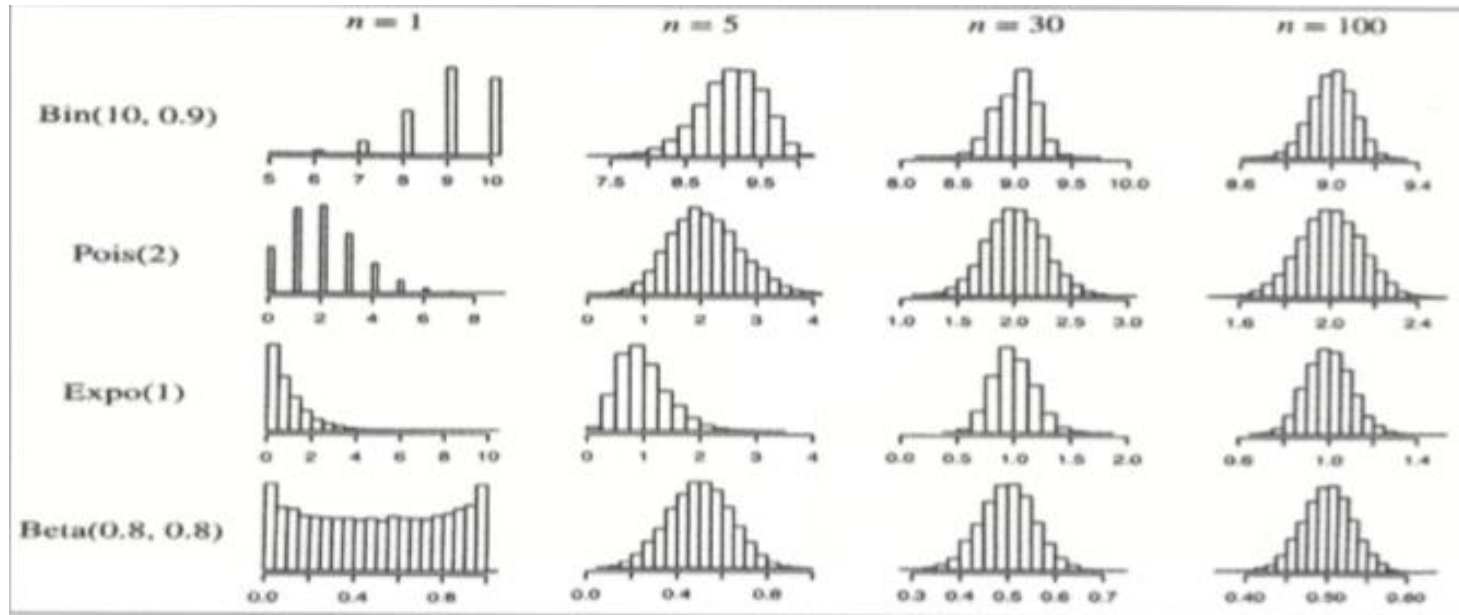
남자의 평균 키는 180cm, 표준편차 10cm이고 평균 몸무게는 78kg, 표준편차 20kg이라 할 때, 명호는 키가 183cm에 몸무게가 73kg이고 한수는 키가 178cm에 몸무게가 78kg이다. 이 때, 키 5cm 차이와 몸무게 5kg 차이 중 어떤 차이가 더 큰가?

- 키 5cm차이의 표준점수?  $\frac{183-180}{10} - \frac{178-180}{10} = \frac{3}{10} - \frac{2}{10} = \frac{1}{2}$
  - 몸무게 5kg차이의 표준점수?  $\frac{78-78}{20} - \frac{73-78}{20} = 0 - \frac{4}{20} = -\frac{1}{4}$
- 키 5cm가 몸무게 5kg보다 표준점수가 크므로 **키 5cm가 몸무게 5kg보다 더 차이가 크다고 볼 수 있음**

## Focus 4. 중심극한정리란 무엇인가?

### ■ 중심극한정리

- 모집단의 분포 형태에 상관없이 얻어진 **표본평균분포는 점차적으로 정규분포에 수렴함**
- **표본의 크기가 커질수록** 이러한 수렴의 속도는 빨라 짐
- **표본평균분포의 평균은 모집단의 평균값과 일치**하여 가는 형태를 보임



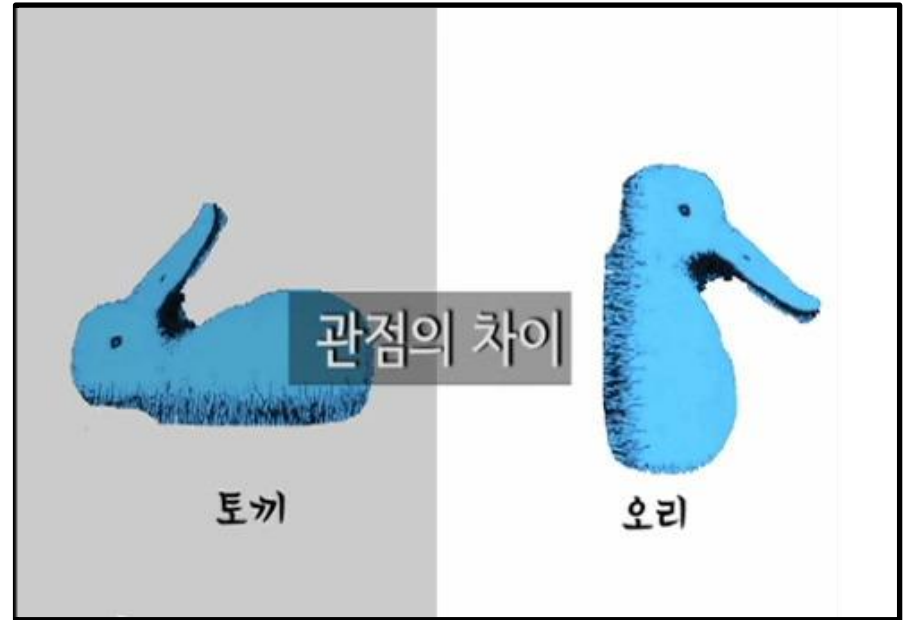
## 1.3 데이터 요약

**정의** ✓ 데이터의 전반적인 내용을 쉽고 빠르게 파악할 수 있도록 하는 방법

**목적** ✓ 데이터에 내재된 정보의 특성을 왜곡하지 않고 정확하게 파악하기 위해 왜곡된 사례와 주의해야하는 데이터를 살펴보고 올바른 데이터 요약 및 표현 방법을 파악하기 위함

- Focus!**
1. 데이터가 나타내는 의미가 왜곡된 경우가 있는가?
  2. 데이터의 왜곡됨을 줄이기 위해 살펴봐야 하는 데이터는 어떤 것인가?
  3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?
  4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## Focus 1. 데이터가 나타내는 의미가 왜곡된 경우가 있는가?



- 같은 그림이 왜 다르게 보일까요?
- **바라보는 시야의 차이!!**  
즉, 데이터를 바라보는 관점에 따라 결과가 왜곡될 수 있음



## Focus 1. 데이터가 나타내는 의미가 왜곡된 경우가 있는가?

### ■ 심슨의 패러독스

여러 하위 집단에서 나타나는 결과와 이들을 결합하여 한 집단으로 놓고 보았을 때의 결과가 다른 현상을 일컫는 용어로 **합계가 부분을 왜곡하거나 부분이 합계를 왜곡**하는 경우를 의미

Ex) A대학교의 특정 학과는 남학생보다 여학생의 합격률이 높다?

### ■ 평균의 함정

평균은 데이터의 전체적인 상태를 나타내는 좋은 방법이지만, 평균만으로 전체를 나타낼 수 없는 경우가 많음. 이런 문제가 발생하는 경우를 의미

Ex) 국회의원 평균 재산이 1,172억원?

## Focus 1. 데이터가 나타내는 의미가 왜곡된 경우가 있는가?

- A대학교의 특정 학과는 남학생보다 여학생의 합격률이 높다?

공학부는 900명 모집, 식품영양학과 100명 모집하는 A대학교에 남학생 1,000명, 여학생 1,000명 지원함 공학부에 남학생 80%합격, 여학생 90%합격하고 식품영양학과에 남학생 10%합격, 여학생 11%합격

### < 공학부\_900명 모집 >

	지원자	합격자	합격률
남학생	900명	720명	80.0%
여학생	200명	180명	90.0%

### < 식품영양학과\_100명 모집 >

	지원자	합격자	합격률
남학생	100명	10명	10.0%
여학생	800명	90명	11.2%

### < 전체 >

	합격자	합격률
남학생	730명	73.0%
여학생	270명	27.0%

학과별 모집단위에서 여학생의 합격률이 높았으므로 전체적으로 볼 때에도 여학생의 합격률이 높아야 하는데 반대로 남학생의 합격률이 높음을 확인 할 수 있다.

➤ **지원자 수를 고려하지 않아 왜곡된 현상 발생!! 심슨의 패러독스!**

# Focus 1. 데이터가 나타내는 의미가 왜곡된 경우가 있는가?

## ■ 국회의원 평균 재산이 1,210억원?

### <국회의원 중 재산 상위 10명과 재산 하위 10명의 재산 공개>

2014년 정기재산변동 사항

19대 국회의원 재산공개 상·하위 10

(단위: 천원)

상위

순위	성명	소속	현재가액	증가액
1	정몽준	새누리당	2,043,043,018	118,138,066
2	안철수	새정치민주연합	156,924,940	-26,180,186
3	김세연	새누리당	98,550,210	10,561,412
4	박덕홍	새누리당	53,903,536	887,731
5	윤상현	새누리당	17,778,640	745,345
6	강석호	새누리당	16,350,424	2,319,501
7	김우성	새누리당	13,744,138	-17,305
8	정의화	새누리당	10,277,204	-2,125,014
9	심윤조	새누리당	9,534,572	-299,957
10	장병완	새정치민주연합	8,248,405	265,285

⋮

⋮

⋮

⋮

⋮

하위

286	김광진	새정치민주연합	92,720	84,268
287	유은혜	새정치민주연합	90,618	11,939
288	박홍근	새정치민주연합	85,110	62,377
289	김한표	새누리당	78,310	123,056
290	오병윤	통합진보당	69,123	51,811
291	김미희	통합진보당	23,962	-16,031
292	김상민	새누리당	-6,149	-73,590
293	심상정	정의당	-60,482	-358,152
294	강동원	새정치민주연합	-70,035	40,114
295	성완종	새누리당	-754,604	-7,851,116

국회의원 20명에 대한 평균 재산 1210억원

국회의원 18명에 대한 평균 재산 214억원

국회의원 16명에 대한 평균 재산 143억원

국회의원 14명에 대한 평균 재산 93억원

### ▶국회의원 평균 재산은 대체 얼마?

몇 명을 제외한 국회의원의 평균 재산이라고 해도 평균값에 큰 변화는 없어야 하는데 국회의원 20명의 평균값과 국회의원 18명일 때 약6배 차이가 나고 16명일 때 9배 차이가 난다는 것을 확인할 수 있다.

➤ **극단값의 영향**으로 **정확한 평균을 알 수가 없는 현상 발생. 평균의 함정!**

## Focus 2. 데이터의 왜곡됨을 줄이기 위해 살펴봐야 하는 데이터는 어떤 것인가?

여러 하위 집단에서 나타나는 결과와 이들을 결합하여 한 집단으로 놓고 보았을 때의 결과가 다른 현상을 일컫는 용어로 **합계가 부분을 왜곡하거나 부분이 합계를 왜곡**하는 경우를 의미

Ex) A대학교의 특정 학과는 남학생보다 여학생의 합격률이 높다?

### ■ 평균의 함정

평균은 데이터의 전체적인 상태를 나타내는 좋은 방법이지만, 평균만으로 전체를 나타낼 수 없는 경우가 많음. 이런 문제가 발생하는 경우를 의미

Ex) 국회의원 평균 재산이 1,172억원?

## Focus 2. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

### 데이터의 중심위치

---

- 평균
  - 산술평균
  - 조화평균
  - 기하평균
  - 절사평균
- 중위수
- 최빈값

### 데이터의 산포도

---

- 분산
- 표준편차
- 사분위수
- 변동계수
- 범위
- 사분위범위

### 데이터의 분포

---

- 왜도
- 첨도

## Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

### ■ 데이터의 중심위치(1/2)

#### ■ 산술평균

- 데이터의 총합을 데이터의 개수로 나눈 값
- 극단값들의 영향을 크게 받음

[예시1] 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

➡ 산술평균 = 5.5

#### ■ 기하평균

- 데이터의 곱에 데이터 수의 제곱근을 취한 값
- 곱셈으로 계산하는 값의 평균을 구할 때 사용

[예시2] 경제성장률 2배, 경제성장률 0.5배

➡ 기하평균 = ? = 평균 경제성장률 1배?

#### ■ 조화평균

- 데이터의 개수를 각 데이터를 분모로 하는 분수의 총합으로 나눈 값
- 시간에 따른 값의 평균 변화율을 구할 때 사용

[예시3] 40km/h, 60km/h, 120km/h

➡ 조화평균 =  $\frac{3}{\frac{1}{40} + \frac{1}{60} + \frac{1}{120}} = 60\text{km/h}$

## Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

### ■ 데이터의 중심위치(2/2)

#### ■ 절사평균

- 크기 순으로 정렬했을 때, 상하위 일정비율의 데이터를 제외한 산술평균
- 극단값들의 영향을 크게 받지 않음

#### ■ 중앙값

- 크기 순으로 정렬했을 때 중앙에 위치하는 값
- 데이터의 분포가 편향되어 있을 때 유용함

#### ■ 최빈값

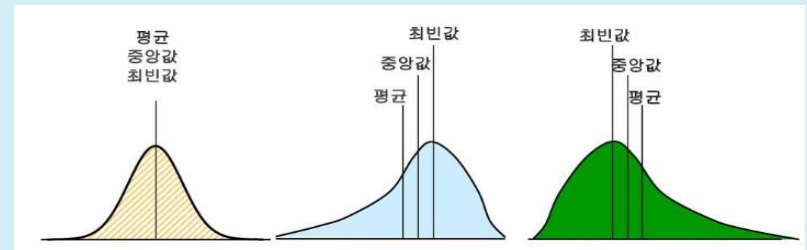
- 데이터 중 빈도가 가장 많은 값
- 평균이나 중앙값을 구하기 어려운 경우 유용함

[예시1] 0.01, 2, 3, 4, 5, 6, 7, 8, 9, 100

➡ 평균 = 5.5, 중위수 = 5.5, 최빈값 = 모든 값

[예시2] 1, 2, 2, 3, 3, 3, 4, 4, 4

➡ 평균 = 3, 중위수 = 3, 최빈값 = 3, 4



# Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

## ■ 데이터의 산포도(1/3)

### ■ 분산

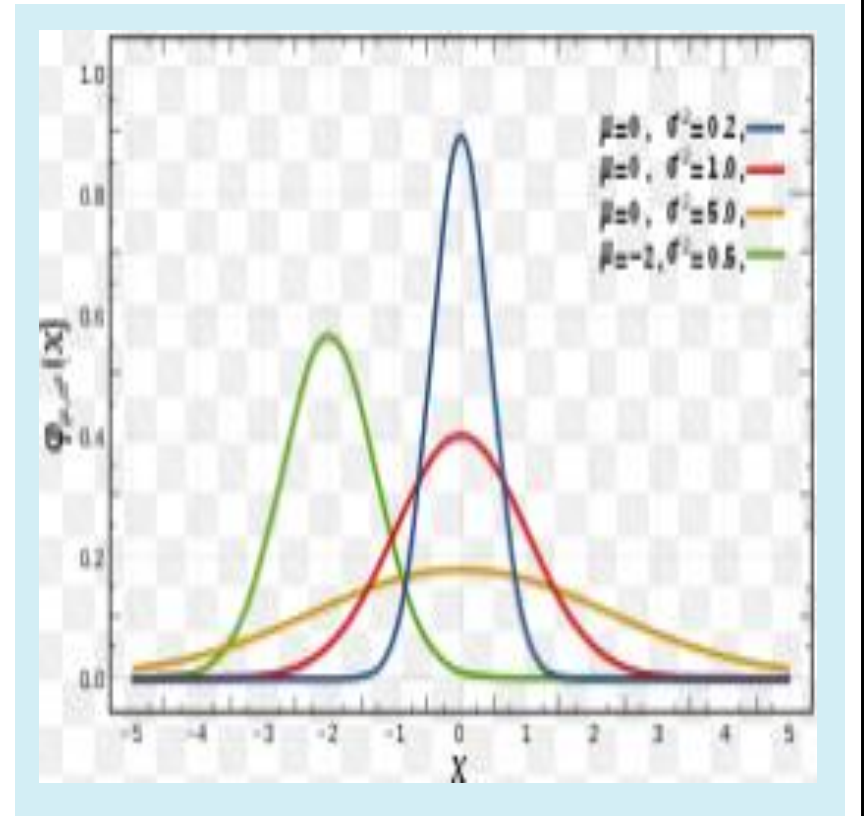
- 평균을 중심으로 데이터의 **흩어진 정도**를 가늠하는 값 (단위 변동 ○)
- 편차(  $d_i = x_i - \mu$  ) 제곱의 평균

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} = \frac{\sum d_i^2}{n}$$

### ■ 표준편차

- 분산의 제곱근 (단위변동 X)
- 항상 0보다 크거나 같음

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}} = \sqrt{\sigma^2}$$



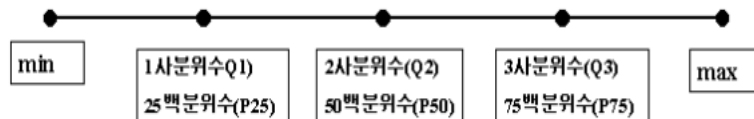


# Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

## ■ 데이터의 산포도(2/3)

### ■ 사분위수

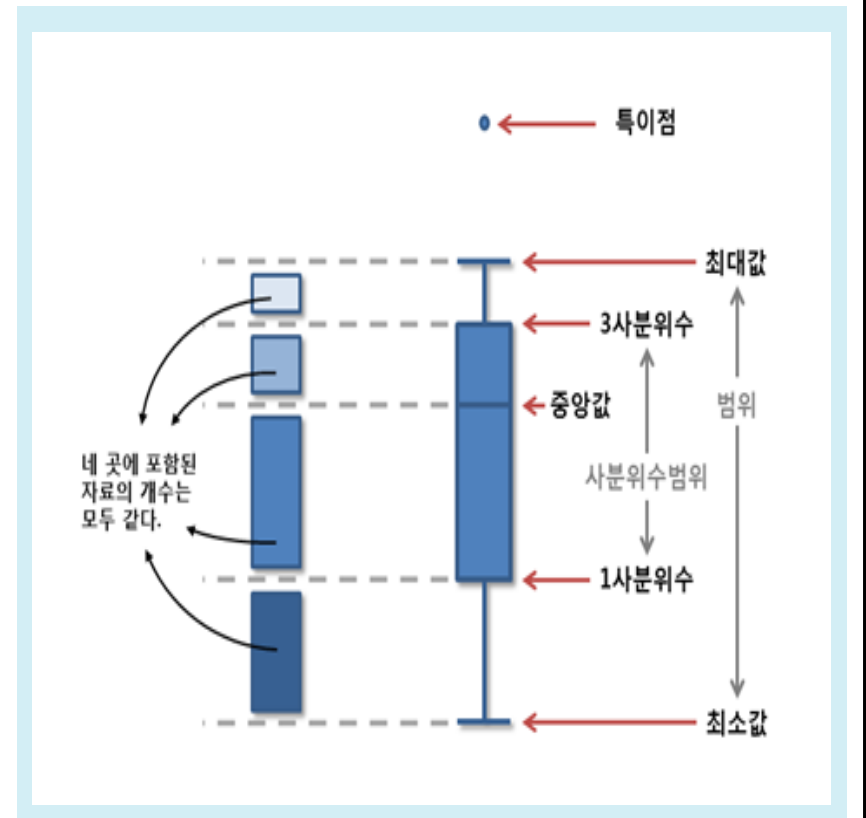
- 전체 데이터를 4등분으로 분할한 값
- 극단값의 영향을 적게 받음



### ■ 변동계수

- 표준편차를 평균에 대한 상대 비율로 표현한 값
- 측정 단위에 영향을 받지 않음

$$cv = \frac{\sigma}{\mu} * 100$$



## Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

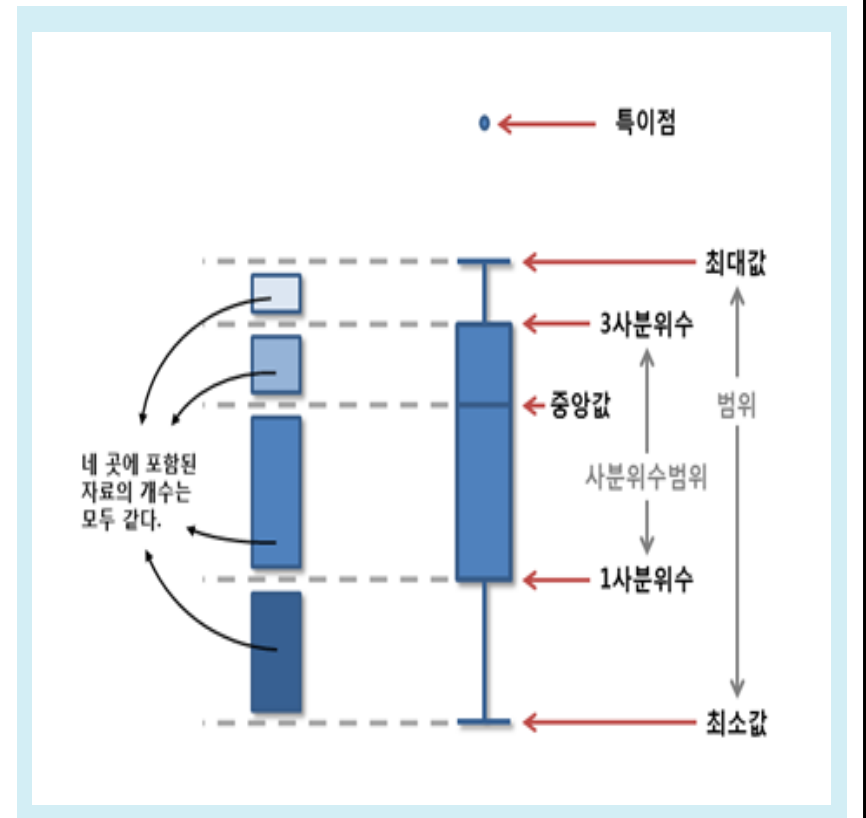
### ■ 데이터의 산포도(3/3)

#### ■ 사분위 범위

- '제3사분위수-제1사분위수'로 계산
- 극단값의 영향을 적게 받음

#### ■ 범위

- '최대값-최소값'으로 계산
- 극단값에 매우 민감하게 영향을 받아 잘 사용하지 않음
- 데이터 분포가 대칭인 경우 사용



## Focus 3. 데이터를 숫자로 요약하고자 할 때 어떤 방법이 있는가?

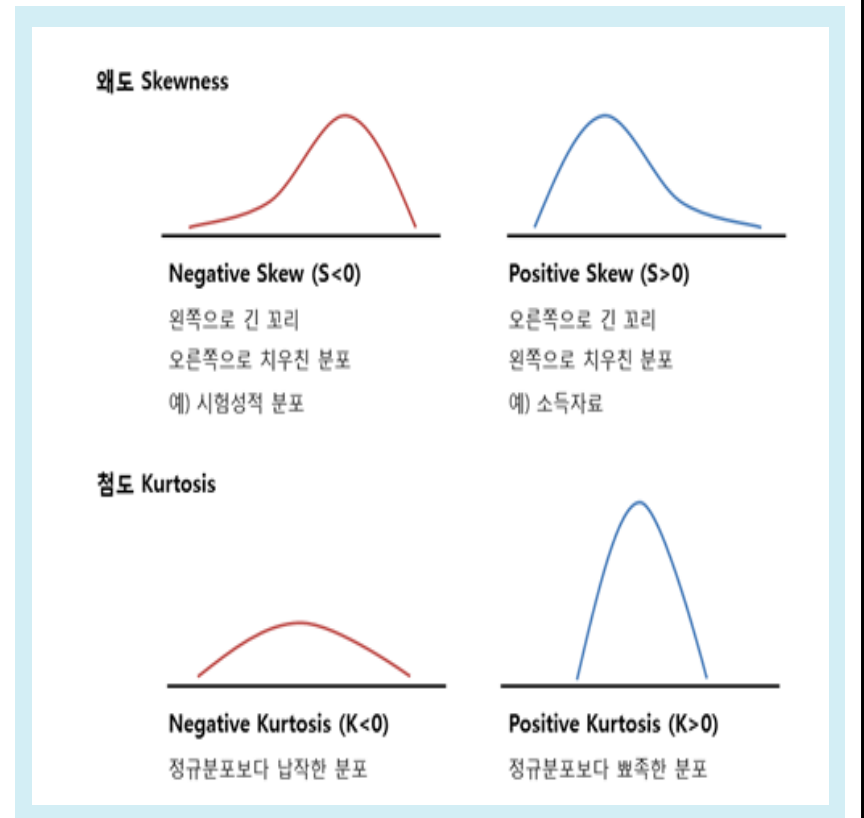
### ■ 데이터의 분포

#### ■ 왜도

- 데이터의 분포에 대한 **비대칭의 정도**
- 왼쪽으로 치우친 경우 : 오른쪽으로 긴 꼬리
- 오른쪽으로 치우친 경우 : 왼쪽으로 긴 꼬리

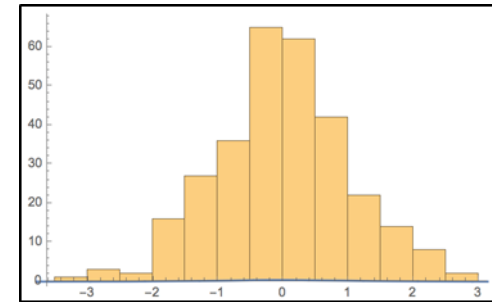
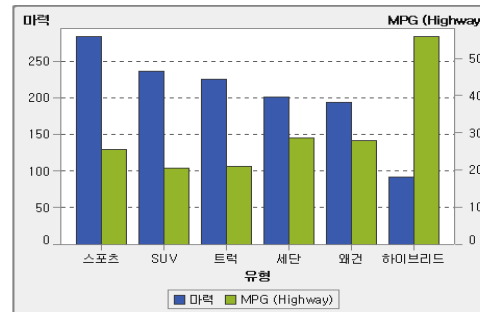
#### ■ 첨도

- 데이터의 분포에 대한 뾰족한 정도
- 꼬리부분의 길이와 중앙부분의 뾰족함에 대한 정보를 제공하는 통계량
- 기준값이 절대적인 값은 아님  
(식에서 3을 빼서 0으로 정의하여 사용하기도 함)



# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

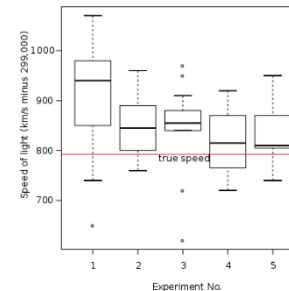
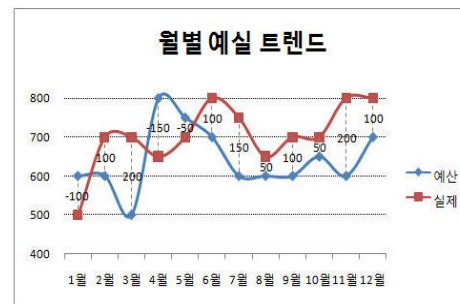
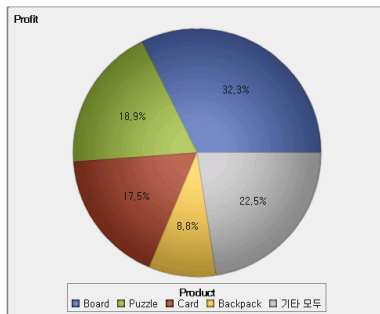
계급	계급값	도수	누적도수	상대도수	누적상대도수
class	value	frequency	cumulative frequency	relative frequency	relative cumulative frequency
0-10	5	5	5	0.167	0.167
10-20	15	8	13	0.267	0.434
20-30	25	4	17	0.133	0.567
30-40	35	2	19	0.067	0.634
40-50	45	5	24	0.167	0.800
50-70	60	6	30	0.200	1.000
합계		30		1	



◎도수분포표

◎막대 그림

◎히스토그램



줄기	잎
6	5 8 2
7	1 4 1 0 9
8	2 7 2 5
9	8 2 3 1

◎원 그림

◎꺾은선 그림

◎상자 그림

◎줄기-잎 그림

## Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

### ■ 도수분포표(1/3)

데이터를 범주(값 또는 구간)에 따라 구분한 후, 각 범주에 속하는 도수(빈도수), 상대도수, 누적상대도수 등을 측정하여 정리한 표

☒ 데이터가 이산형인 경우

☒ 데이터가 연속형인 경우

# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## ■ 도수분포표(2/3)

☑ 데이터가 이산형인 경우

〈동전을 각각 4번씩 던져 앞면이 나온 횟수 데이터〉

학생번호	1	2	3	4	5	...	22	23	24	25
앞면횟수	2	2	0	1	1	...	2	1	2	2

### ※작성 방법

- ① 범주(값 또는 구간)에 대한 리스트 작성
- ② 각 범주별 도수(빈도) 작성
- ③ 각 범주별 상대도수 계산 (상대도수 = 도수 / 전체도수)
- ④ 각 범주별 누적 상대도수 계산

횟수	도수	상대도수	누적 상대도수
0	1	0.04	0.04
1	7	0.28	0.32
2	10	0.40	0.72
3	6	0.24	0.96
4	1	0.04	1.00
전체	25	1.00	

# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## ■ 도수분포표(3/3)

☑ 데이터가 연속형인 경우

〈25개 그룹별 학생들의 통계학과목 평균 데이터〉

그룹번호	1	2	3	...	23	24	25
앞면횟수	71.2	77.5	50.7	...	87.8	63.1	75.0

### ※작성 방법

- ① 범위(최대값 - 최소값)를 계산 → 범위 :  $98 - 50 = 48$
- ② 적당한 범주 (값 또는 구간) 개수를 결정 → 6개로 지정
- ③ 범주 폭(범위/범주 개수)을 계산 → 구간폭 :  $48/6 = 8$
- ④ 계산된 범주 폭에 따라 범주를 설정 후 도수분포표 작성

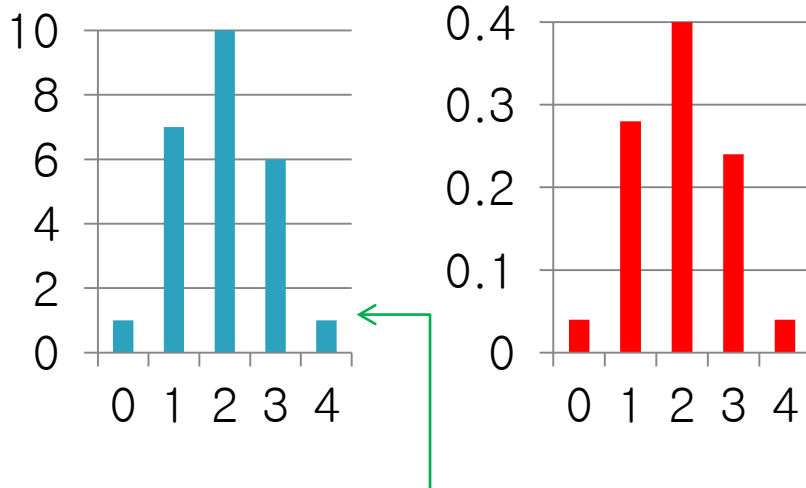
계급 구간	도수	상대도수	누적 상대도수
50이상 ~ 58미만	2	0.08	0.08
58이상 ~ 66미만	3	0.12	0.20
66이상 ~ 74미만	5	0.20	0.40
74이상 ~ 82미만	6	0.24	0.64
82이상 ~ 90미만	5	0.20	0.84
90이상	4	0.16	1.00
전체	25	1.00	

# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## ■ 막대그림

데이터가 **이산형**일 때, 각 범주(값 또는 구간)가 가지는 도수 또는 상대도수를 같은 폭의 막대 모양으로 나타낸 그림으로 **막대는 서로 떨어져 있게 그려야 함**

〈동전을 각각 4번씩 던져 앞면이 나온 횟수 데이터〉



횟수	도수	상대도수	누적 상대도수
0	1	0.04	0.04
1	7	0.28	0.32
2	10	0.40	0.72
3	6	0.24	0.96
4	1	0.04	1.00
전체	25	1.00	

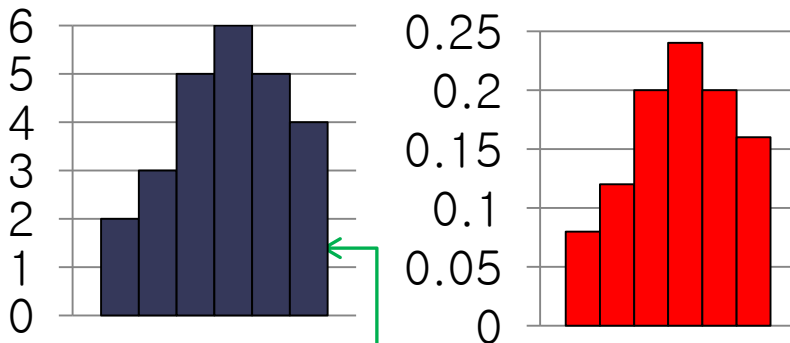


# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## ■ 히스토그램

데이터가 **연속형**일 때, 각 범주(값 또는 구간)가 가지는 도수 또는 상대도수를 같은 폭의 막대 모양으로 나타낸 그림으로 **막대는 사이가 이어지도록 그려야 함**

〈학생들의 통계학과목 점수 데이터〉



계급 구간	도수	상대도수	누적 상대도수
50이상 ~ 58미만	2	0.08	0.08
58이상 ~ 66미만	3	0.12	0.20
66이상 ~ 74미만	5	0.20	0.40
74이상 ~ 82미만	6	0.24	0.64
82이상 ~ 90미만	5	0.20	0.84
90이상	4	0.16	1.00
전체	25	1.00	

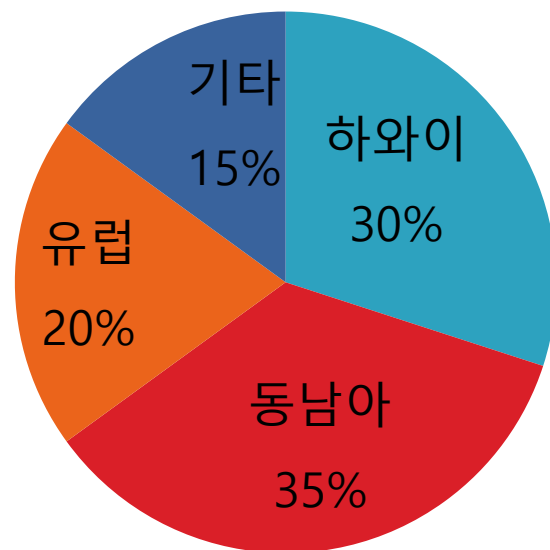
## Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

### ■ 원그림

범주형 데이터의 상대도수와 비례하도록 원의 조각을 나누어 데이터를 표기

〈선호하는 신혼 여행지에 대한 설문 결과〉

신혼 여행지	응답수	상대도수
하와이	12	0.30
동남아	14	0.35
유럽	8	0.20
기타	6	0.15



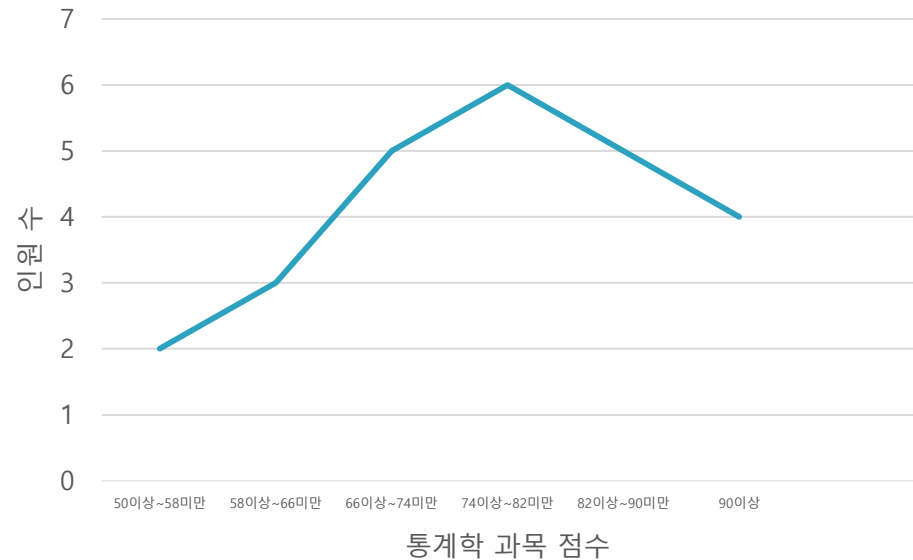
# Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

## ■ 꺾은선 그림

도수분포표와 같은 내용을 **선으로 이어 표현**하는 그림

〈25명의 학생들의 통계학 과목 점수 데이터〉

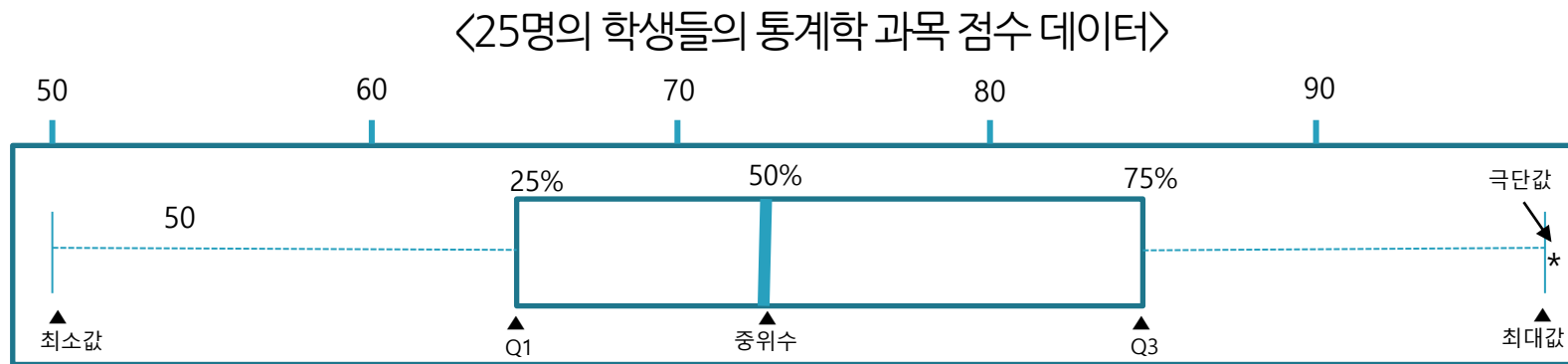
계급 구간	도수
50이상~58미만	2
58이상~66미만	3
66이상~74미만	5
74이상~82미만	6
82이상~90미만	5
90이상	4
전체	25



## Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

### ■ 상자 그림

다섯 가지의 숫자 (최소값, 제1사분위수, 중위수, 제3사분위수, 최대값)로 데이터의 분포를 나타낸 그림으로 최소값보다 작거나 최대값보다 큰 값인 **극단값**은 '\*' 모양으로 표시



-최소값 : 제1사분위수(Q1) - 1.5\*사분위범위(IQR)

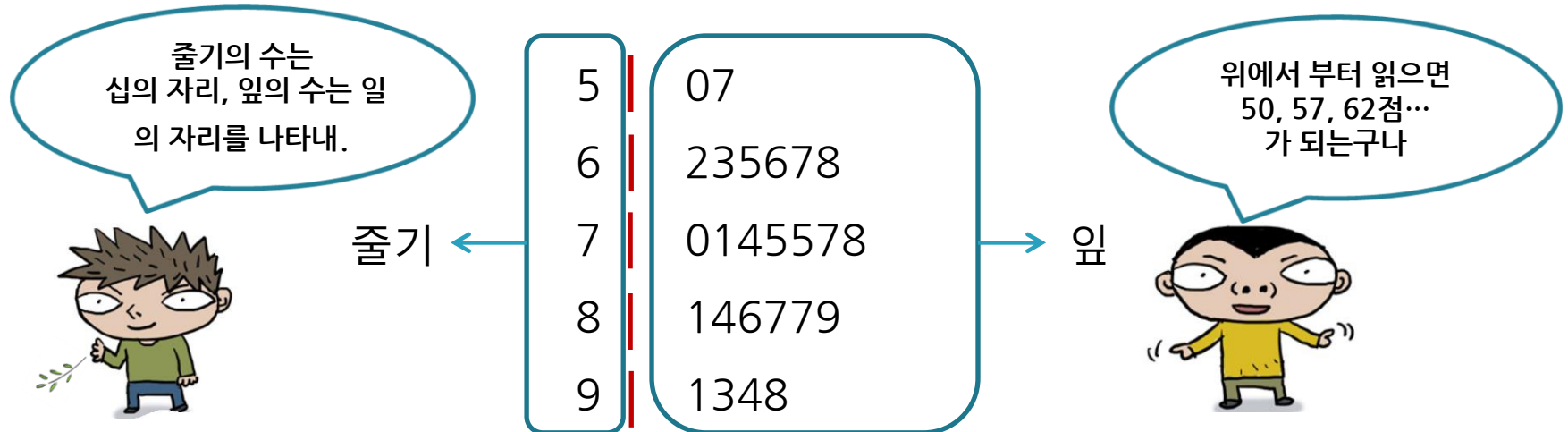
-최대값 : 제3사분위수(Q3) + 1.5\*사분위범위(IQR)

## Focus 4. 데이터를 시각적으로 표현하고자 할 때 어떤 방법이 있는가?

### ■ 줄기-잎 그림

측정한 값의 자리 수를 적절히 구분하여 앞자리를 줄기라 하고 뒷자리를 잎이라 하여  
같은 줄기를 가지는 잎의 도수분포를 나타낸 그림

〈25명의 학생들의 통계학 과목 점수 데이터〉



## 1.4. 검정 통계

**정의** ✓ 기업의 비즈니스 업무, 경제 사회에서 발생하는 문제 등을 통계적으로 입증하기 위해 분석 목적 시나리오에 따른 **가설을 세우고 통계적으로 맞는 지 틀린 지를 확인하는 방법**

**목적** ✓ 정의된 가설을 검정할 때 알아야하는 용어 및 가설의 정의와 방법을 이해하고 검정 형태에 따른 구분과 검정 오류란 무엇인지를 파악하기 위함

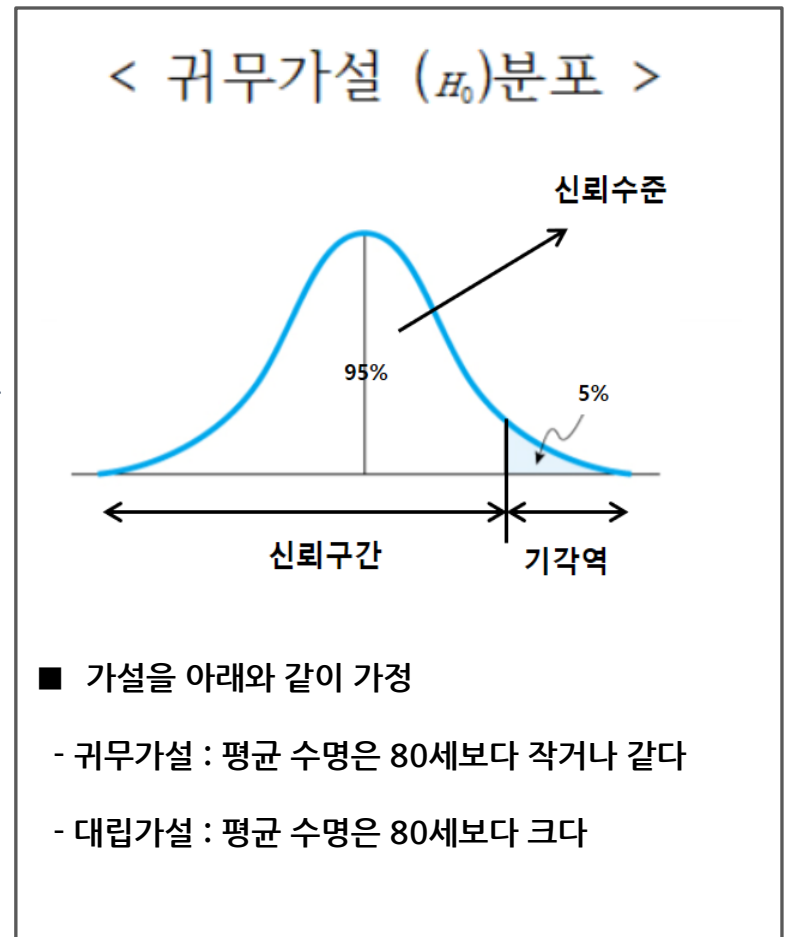
- Focus!**
1. 검정 통계를 이해하기 위해 알아야 하는 **용어**는 무엇인가?
  2. 가설이란 무엇이고 어떻게 작성하는가?
  3. 어떤 방식으로 가설을 검정하는가?
  4. 검정형태는 어떻게 구분되는가?
  5. 검정 오류란 무엇인가?

# Focus 1. 검정 통계를 이해하기 위해 알아야 하는 용어는 무엇인가?

- 귀무가설 : 일반적으로 인정하는 주장
- 대립가설 : 통계적 근거를 통해 입증하려는 주장
- 신뢰구간 : 귀무가설이 채택될 것으로 추정하는 구간
- 신뢰수준( $1-\alpha$ ) : 검정을 반복할 때 귀무가설이 채택될 확률

예) 90% 신뢰수준, 95% 신뢰수준, 99% 신뢰수준

- 기각역 : 귀무가설이 기각될 것으로 추정하는 구간



오른쪽 단측검정에 대한 정규분포

# Focus 1. 검정 통계를 이해하기 위해 알아야 하는 용어는 무엇인가?

- 유의수준( $\alpha$ ) : 검정을 반복할 때 귀무가설이 기각될 확률

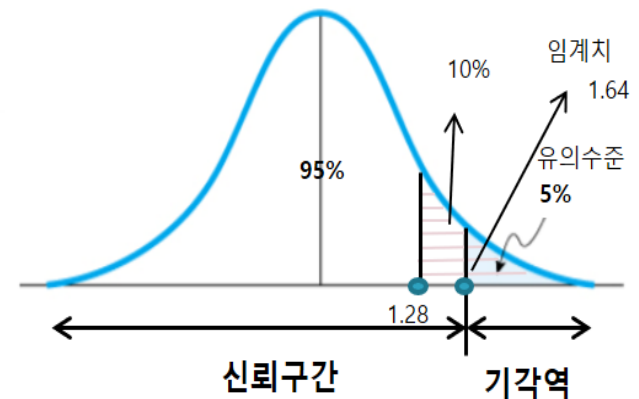
예) 유의수준 0.10, 유의수준 0.05, 유의수준 0.01

- 임계치 : 신뢰구간과 기각역의 경계선에 대응되는 값

- 유의확률(p-값) : 검정하고자 하는 데이터로 계산한 귀무가설이 기각될 확률

- 검정 통계량 : 유의확률에 대응하는 값

## < 귀무가설 ( $H_0$ ) 분포 >

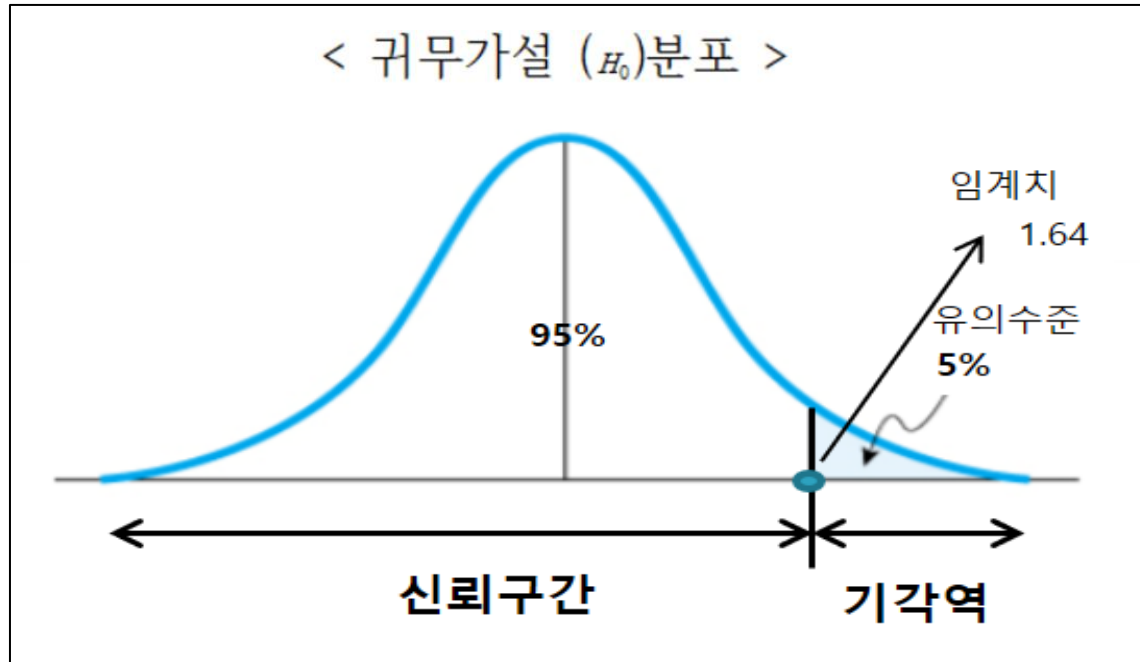


### ■ 분석결과를 아래와 같이 가정

- 유의확률 : 10%
- 검정 통계량 : 1.28



# Focus 1. 검정 통계를 이해하기 위해 알아야 하는 용어는 무엇인가?



✓ 임계치와 검정통계량을 비교하여,

임계치  $\geq$  검정통계량이 성립되면 귀무가설 채택

임계치  $<$  검정통계량이 성립되면 대립가설 채택

✓ 유의수준과 유의확률(p-값)을 비교하여,

유의수준  $\leq$  유의확률이 성립하면 귀무가설 채택

유의수준  $>$  유의확률이 성립하면 대립가설 채택

➡ 분석 방법에 따른 임계치를 구하기 어려워, 일반적으로 유의수준을 활용하여 판단함

## Focus 2. 가설이란 무엇이고 어떻게 작성하는가?

### 귀무가설( $H_0$ )

- 일반적으로 인정하는 주장
- 예) 평균이 차이가 없다  
젊음과 건강은 상관이 없다  
영향력이 없다  
\*\*감기약은 효과가 없다  
작거나 같다/ 크거나 같다

### 대립가설( $H_1$ )

- 통계적 근거를 통해 입증하려는 주장
- 예) 평균이 차이가 있다  
젊음과 건강은 상관이 있다  
영향력이 있다  
\*\*감기약은 효과가 있다  
작다/ 크다



가설이란, 연구 문제에 대한 주장이나 서술을 의미함

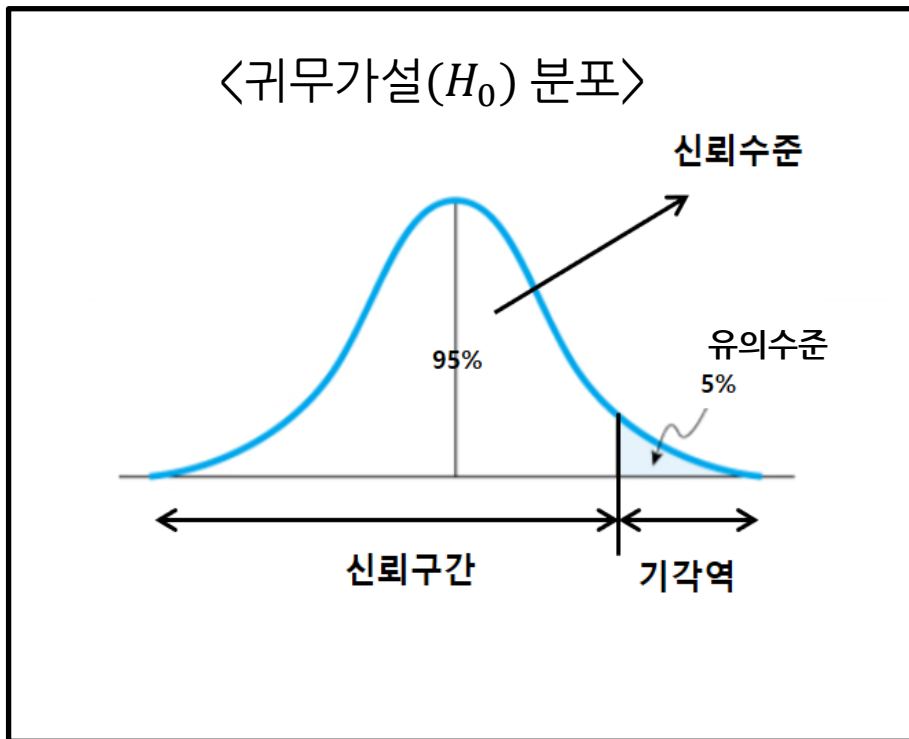
## Focus 3. 어떤 방식으로 가설을 검정하는가(1/2)

- 1 가설 설정 ( 귀무가설 ( $H_0$ ), 대립가설 ( $H_1$ ) )
- 2 가설에 적합한 분석 방법 (평균 차이 검정, 분산 분석, 상관관계 분석, 회귀 분석 등) 선택
- 3 유의수준 ( $\alpha$ ) 설정 ( 예 : 0.10, 0.05, 0.01 )
- 4 유의확률 (p-값) 계산 혹은 검정 통계량 계산
- 5 가설 판정 (  $p\text{-값} < \alpha$ ,  $|t\text{-값}| > |t_{\text{critical}}|$  이면 귀무가설 기각, 대립가설 채택 )

## Focus 3. 어떤 방식으로 가설을 검정하는가(2/2)

- 1 귀무가설( $H_0$ ) : A반 중간고사 평균과 B반 중간고사 평균은 차이가 없다  
대립가설 ( $H_1$ ) : A반 중간고사 평균과 B반 중간고사 평균은 차이가 있다
- 2 평균 차이 검정
- 3 유의수준( $\alpha$ ) = 0.05
- 4 유의확률(p-값) = 0.02
- 5  $0.02 < 0.05$ 이므로 귀무가설을 기각하여, 반 평균 차이가 있다(대립가설 채택)고 말할 수 있음

## Focus 4. 검정 형태는 어떻게 구분되는가?(1/2)



① 유의수준의 의미가 무엇인가?

검정을 반복할 때 귀무가설이 기각될 확률  
즉, 검정통계량 값이 기각역에 해당하는 경우가 귀무가설을 기각하는 경우를 의미

② 분포의 중심은 무엇을 의미하는가?

평균

③ 유의수준 영역의 위치는 어떤 의미를 담고있겠는가?

위 그림에서는 기각역이 평균보다 오른쪽에 위치하므로 대립가설이 '평균보다 크다'임을 알 수 있음

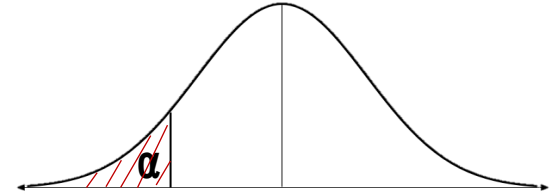
➡ 기각역의 위치는 대립가설의 방향성을 나타냄

## Focus 4. 검정 형태는 어떻게 구분되는가?(2/2)

- **왼쪽  
단측검정**

기각역이 **왼쪽**으로 구성되는 검정

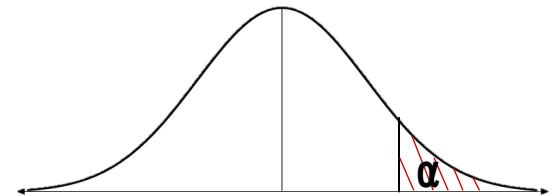
예)  $H_0$  : 우리 회사 평균 나이는 40세보다  
많거나 같다



- **오른쪽  
단측검정**

기각역이 **오른쪽**으로 구성되는 검정

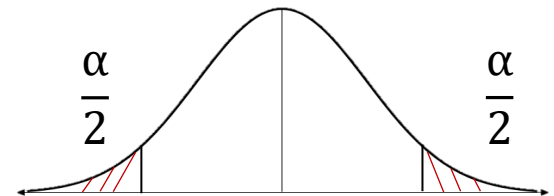
예)  $H_0$  : 우리 회사 평균 나이는 40세보다  
적거나 같다



- **양측검정**

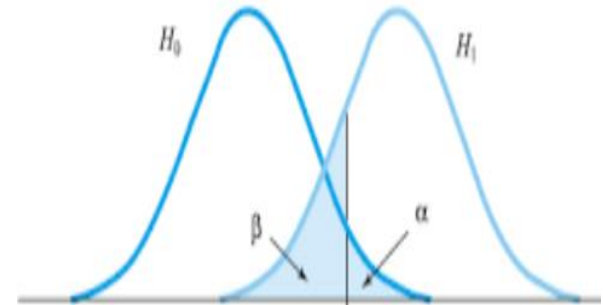
기각역이 **양쪽**으로 구성되는 검정

예)  $H_0$  : 우리 회사 평균 나이는 40세다



## Focus 5. 검정 오류란 무엇인가?

실제 \ 결정	귀무가설 참	귀무가설 거짓
귀무가설 채택	옳은 결정( $1-\alpha$ )	제2종 오류( $\beta$ )
귀무가설 기각	제1종 오류( $\alpha$ )	옳은 결정( $1-\beta$ )



- 제1종 오류 : 귀무가설이 맞는데, 귀무가설이 틀리다고 잘못 판단하여 기각함(대립가설을 채택)
- 제2종 오류 : 귀무가설이 틀린데, 귀무가설이 맞다고 잘못 판단하여 채택함(대립가설을 기각)

☑ 사례. 재판에서의 경우

$H_0$  : 피고인은 살인을 하지 않았다

$H_1$  : 피고인은 살인을 했다

➡ 제1종 오류 : 피고인은 살인을 하지 않았는데 유죄(살인을 했다)로 판결이 난 경우

제2종 오류 : 피고인은 살인을 했는데 무죄(살인하지 않았다)로 판결이 난 경우

# 목차

---

## 1. 기초 통계 이론

---

## 2. 기초 통계 분석

---

2.1 빈도 분석 / 기술 통계 분석

2.2 교차 분석

2.3 평균 차이 검정

2.4 분산 분석

2.5 상관관계 분석

2.6 회귀 분석

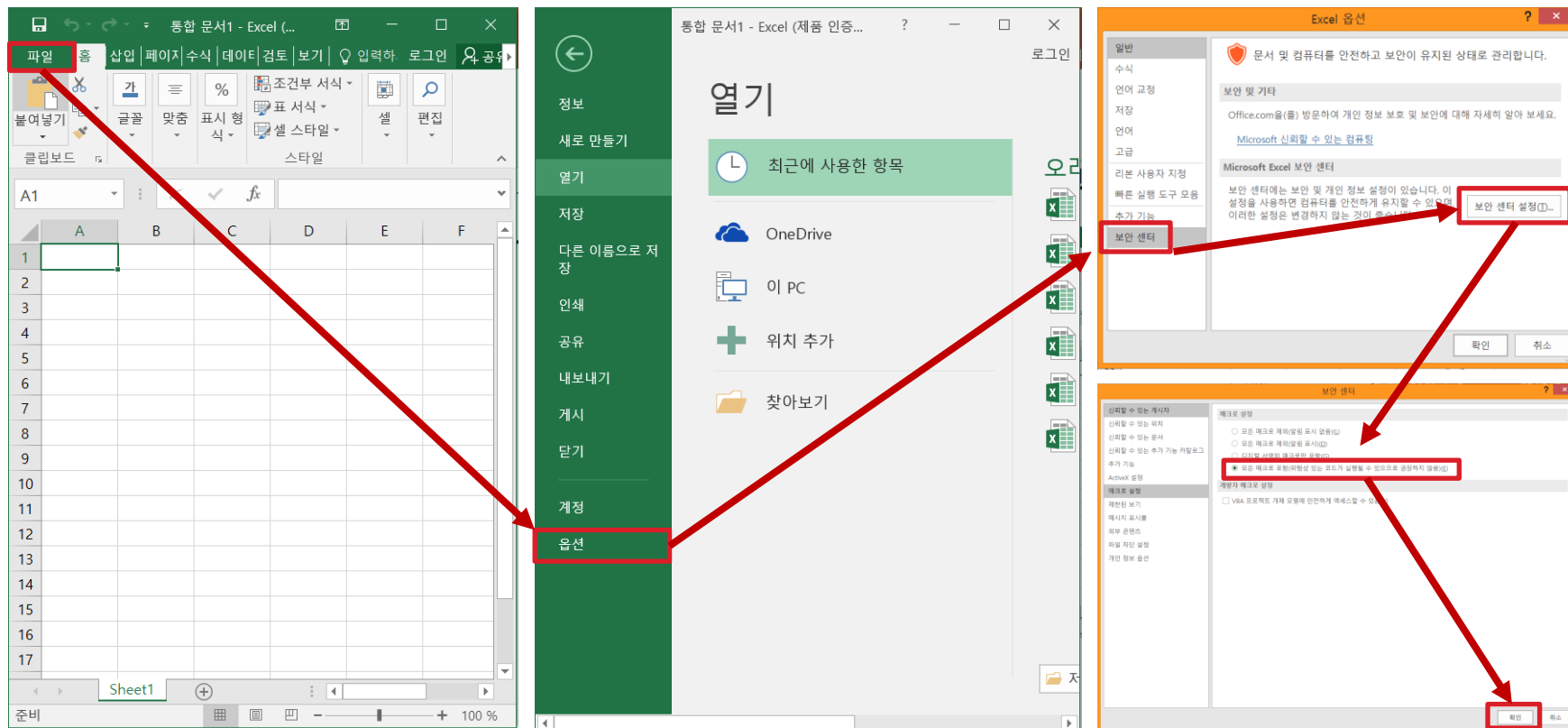


# [기초 통계 분석기법] 들어가기에 앞서...

**학습목표** ✓ 기초 통계 분석기법들에 대한 이론과 실습 방법을 파악해 본다

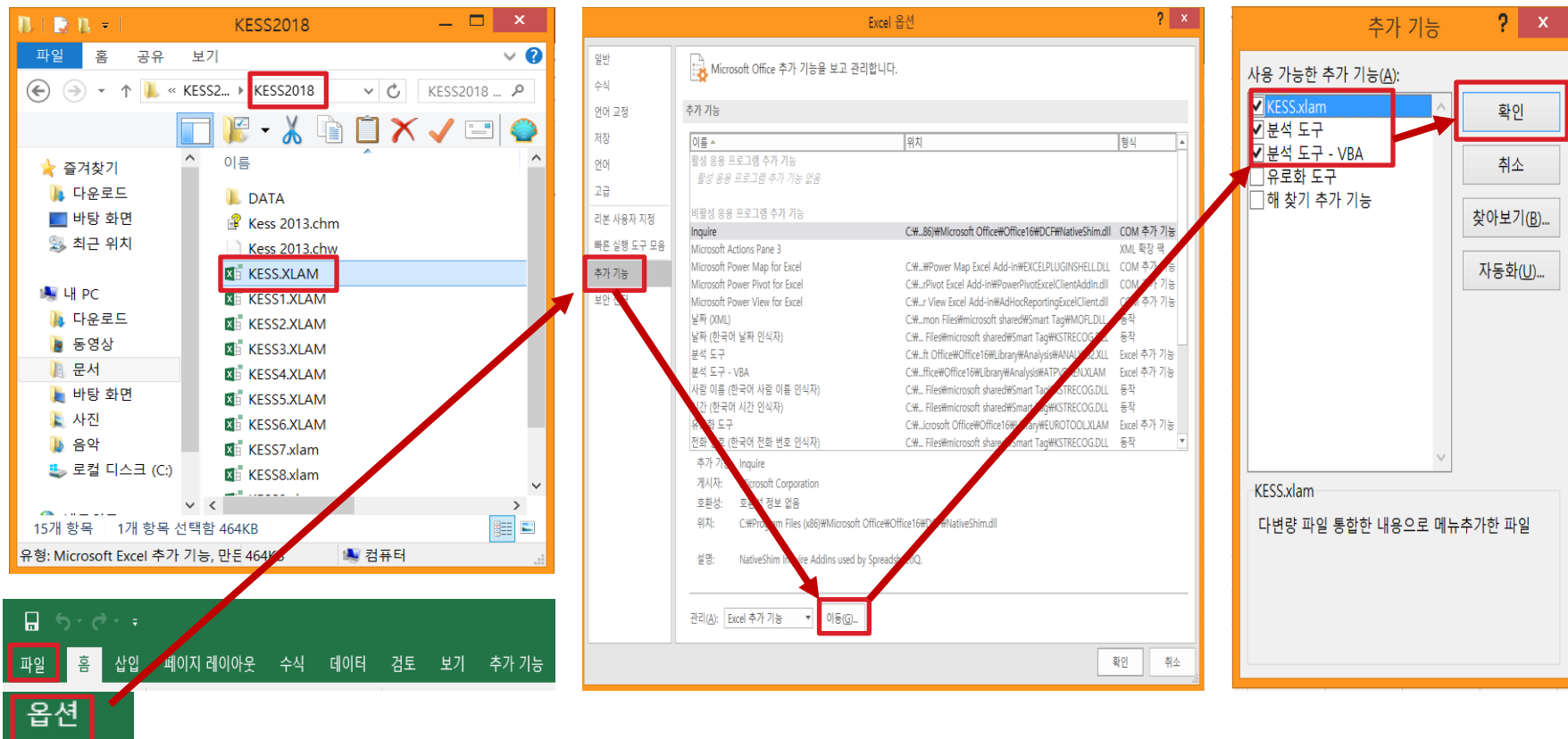
**학습내용** ✓ 기초 통계 분석에는 어떤 **기법**들이 있는가?  
✓ **엑셀**을 통해 각각의 분석 기법을 **실습**하고 **결과**를 **해석**해보자

# 엑셀 매크로 보안 변경



[파일] 탭 → [옵션] 클릭 → [보안 센터] 클릭 → [보안 센터 설정] 클릭  
→ [매크로 설정] 클릭 → [모든 매크로 포함] 클릭

# KESS 분석 메뉴 생성

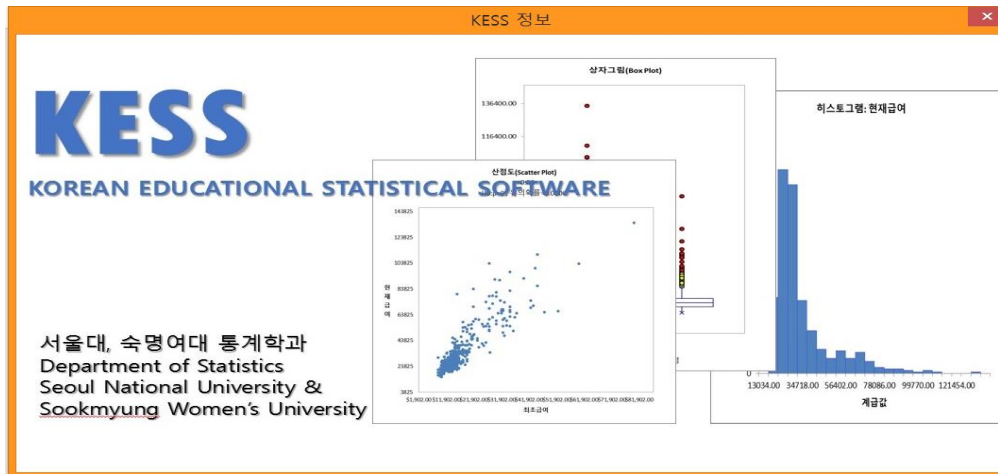


[KESS2018]폴더 → [KESS.XLAM] 파일 → [파일] 탭 → [옵션] 탭 → [추가 기능] 클릭 → [이동] 클릭 →  
 [KESS.xlam], [분석도구], [분석 도구-VBA] 클릭 → 확인(만약, KESS.xlam이 없을 경우 [찾아보기]를 통해 찾기)  
 확인사항 : [데이터] 탭 → [분석] 그룹 → [데이터 분석]과 [추가 기능] 탭 → [메뉴 명령] 그룹 → [통계분석]이 생성필요

## 2.1 빈도 분석 / 기술통계 분석

**정의** ✓ 데이터의 전반적인 내용을 쉽고 빠르게 파악하기 위해 정리하는 방법

**목적** ✓ 빈도 분석과 기술통계 분석 결과를 보고 어떻게 결과를 정리해야 효율적일지 작성 방법을 파악하기 위함



# KESS 분석 메뉴 생성

number	x1	x2	x3
1	2	2	3
2	4	3	4
3	3	4	3
4	4	4	3
5	4	4	4

-공백이 없어야 함(0으로 값 입력 필요)

[추가 기능]탭 → [메뉴 명령]그룹 →

[통계분석] → [설문분석] →

[빈도 분석]선택 → [빈도 분석] 옵션 창

에서 [선택변수] 리스트 중 빈도 분석을

할 변수 선택 후 화살표 (→) 를 통하여

[분석변수] 리스트로 이동 → 확인

The screenshot shows the KESS software interface with the following elements highlighted:

- Menu Path:** The '추가 기능' (Additional Functions) tab is selected. The '통계분석' (Statistical Analysis) menu is open, showing '설문분석' (Survey Analysis) and '빈도 분석' (Frequency Analysis) selected.
- Excel Data:** An Excel spreadsheet is visible in the background with columns labeled 'number', 'x1', 'x2', 'x3', 'x4', 'x5' and rows of data.
- Frequency Analysis Dialog Box:** A dialog box titled '빈도 분석' (Frequency Analysis) is open. It has two lists: '선택변수' (Selected Variables) and '분석변수' (Analysis Variables). The 'number' variable is in the '선택변수' list, and the 'x1', 'x2', 'x3', 'x4', 'x5' variables are in the '분석변수' list. The '확인' (OK) button is highlighted.

# KESS 활용한 분석 결과

## 빈도분석결과

×1

데이터 범주	빈도수	비율
1	1	5%
2	4	20%
3	5	25%
4	10	50%
총합계	20	100%

×2

데이터 범주	빈도수	비율
1	1	5%
2	4	20%
3	7	35%
4	8	40%
총합계	20	100%

×3

데이터 범주	빈도수	비율
1	1	5%
2	2	10%
3	10	50%
4	7	35%
총합계	20	100%

×4

데이터 범주	빈도수	비율
2	5	25%
3	7	35%
4	8	40%
총합계	20	100%

×5

데이터 범주	빈도수	비율
1	2	10%
2	10	50%
3	6	30%
4	2	10%
총합계	20	100%

# 함수를 활용한 분석 방법 및 결과

## <데이터 주의사항>

number	x1	x2	x3
1	2	2	3
2	4	3	4
3	3	4	3
4	4	4	3
5	4	4	4

- 공백이 없어야 함 (0으로 값 입력 필요)

## <결과>

	x1	x2	x3	x4	x5
1	1	1	1	0	2
2	4	4	2	5	10
3	5	7	10	7	6
4	10	8	7	8	2

= COUNTIF( range, criteria )

: range에 포함되는 데이터 중 criteria 값을 갖는 데이터의 개수 출력

	x1	x2	x3	x4	x5
1	=COUNTIF(B\$2:B\$21,\$H4)	=COUNTIF(C\$2:C\$21,\$H4)	=COUNTIF(D\$2:D\$21,\$H4)	=COUNTIF(E\$2:E\$21,\$H4)	=COUNTIF(F\$2:F\$21,\$H4)
2	=COUNTIF(B\$2:B\$21,\$H5)	=COUNTIF(C\$2:C\$21,\$H5)	=COUNTIF(D\$2:D\$21,\$H5)	=COUNTIF(E\$2:E\$21,\$H5)	=COUNTIF(F\$2:F\$21,\$H5)
3	=COUNTIF(B\$2:B\$21,\$H6)	=COUNTIF(C\$2:C\$21,\$H6)	=COUNTIF(D\$2:D\$21,\$H6)	=COUNTIF(E\$2:E\$21,\$H6)	=COUNTIF(F\$2:F\$21,\$H6)
4	=COUNTIF(B\$2:B\$21,\$H7)	=COUNTIF(C\$2:C\$21,\$H7)	=COUNTIF(D\$2:D\$21,\$H7)	=COUNTIF(E\$2:E\$21,\$H7)	=COUNTIF(F\$2:F\$21,\$H7)

# mpg 데이터

## <데이터 설명>

- 미국의 자동차 연비 측정 데이터 (1999~2008)
- R프로그램의 ggplot2 패키지 제공 데이터로써 11개의 변수와 234개 차종의 관측값
  - trans(변속기 유형), model(차종), year(연식), class(차 분류) 등이 기록되어 있음

## <목적>

변속기 유형(trans)을 크게 2가지로 분류를 한 후 각각에 대한 빈도를 분석하고자 함

## <분석 과정>

- ① 변속기 유형(trans)의 값 확인
- ② 변속기 유형(trans)의 값을 자동(manual), 수동(auto)으로만 구분하는 변수(re\_trans) 생성
- ③ re\_trans에 대한 빈도 분석



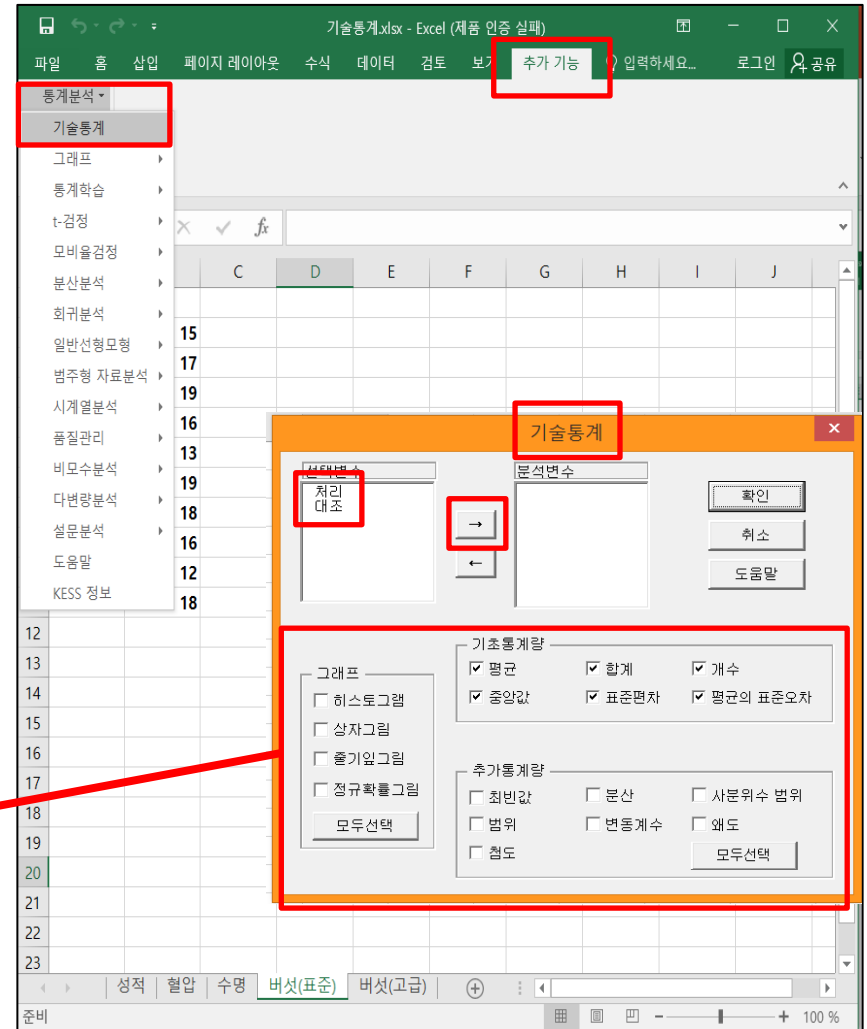
# KESS를 활용한 분석 방법

[추가 기능]탭 → [메뉴 명령]그룹 →  
[통계분석]→[기술통계]선택 →  
[기술통계]옵션 창에서 [선택변수]리스트 중  
기술통계량을 계산 할 변수를 선택 후 화살  
표(→)를 통하여 [분석변수] 리스트로 이동  
→값들의 기준 선택 및 추출할 통계량과 그  
래프의 종류를 선택 →확인 클릭

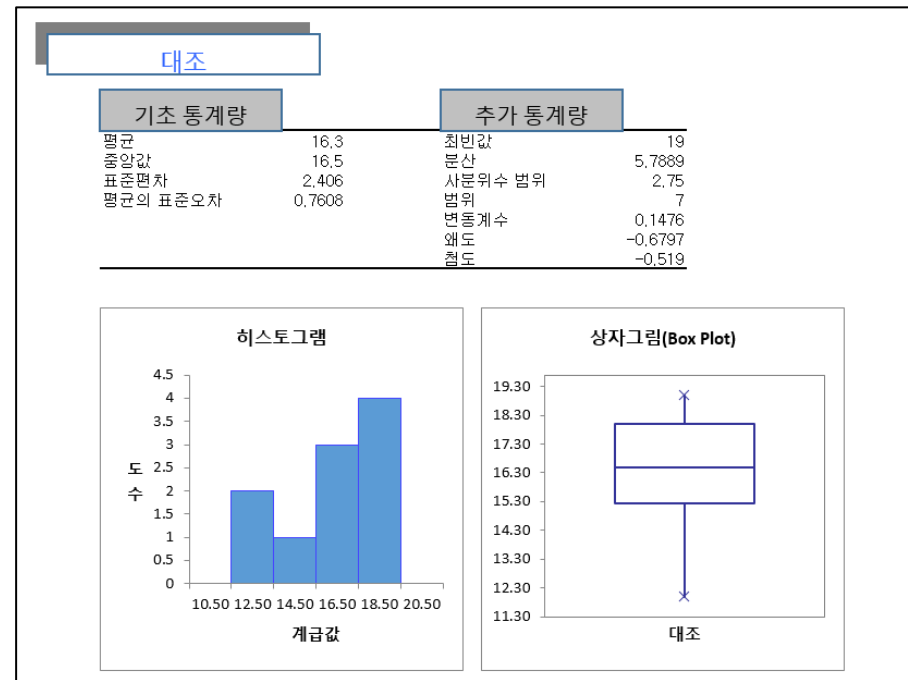
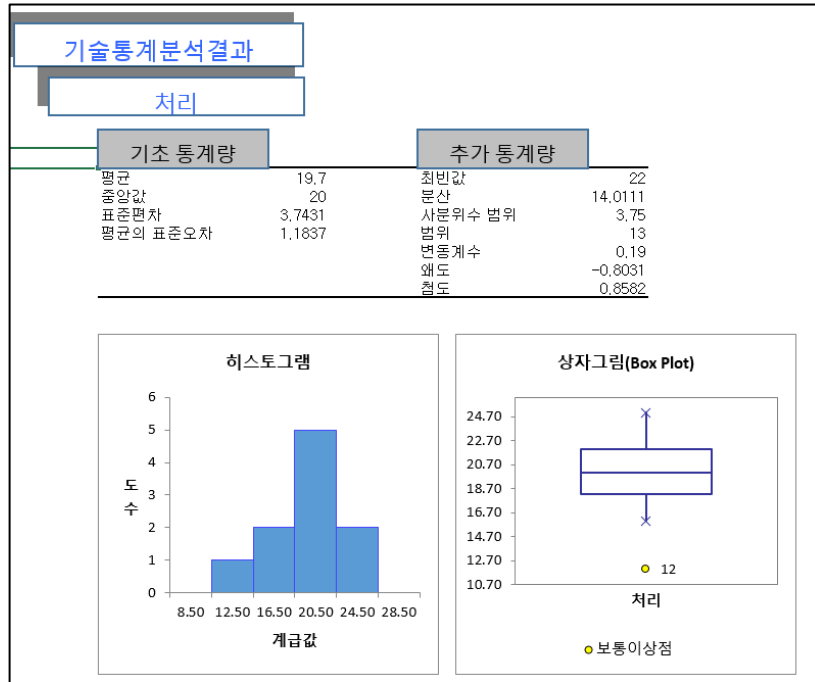
그래프 —  
☒ 히스토그램  
☒ 상자그림  
☐ 줄기잎그림  
☐ 정규확률그림  
모두선택

기초통계량 —  
☒ 평균  
☒ 중값  
☒ 표준편차  
☐ 합계  
☐ 개수  
☒ 평균의 표준오차

추가통계량 —  
☒ 최빈값  
☒ 범위  
☒ 첨도  
☒ 분산  
☒ 변동계수  
☒ 사분위수 범위  
☒ 왜도  
선택취소

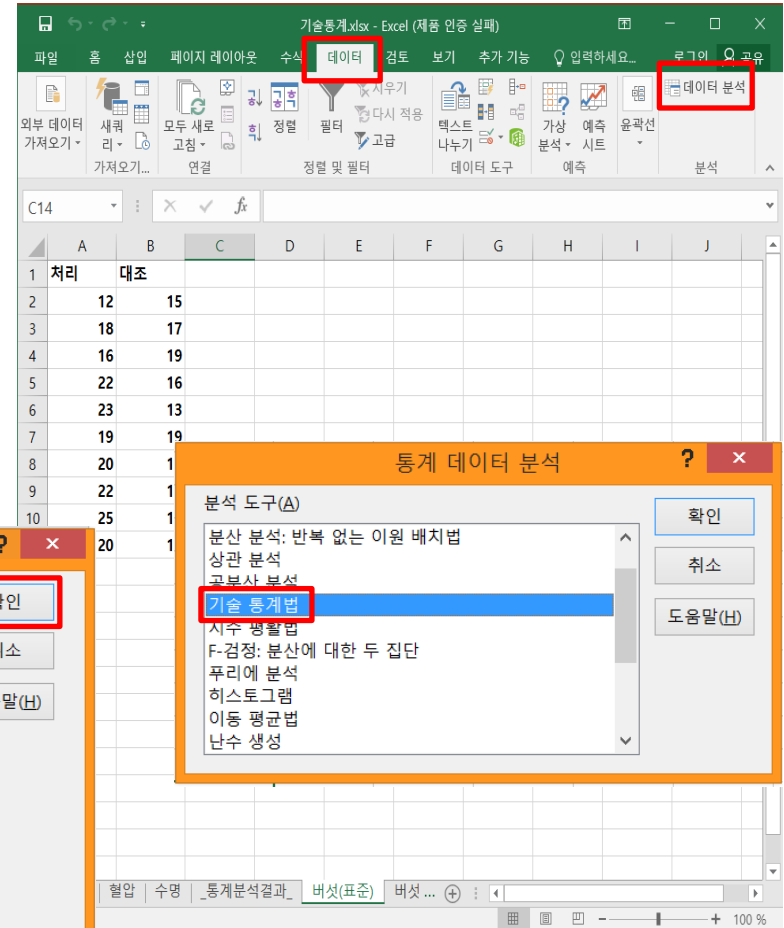
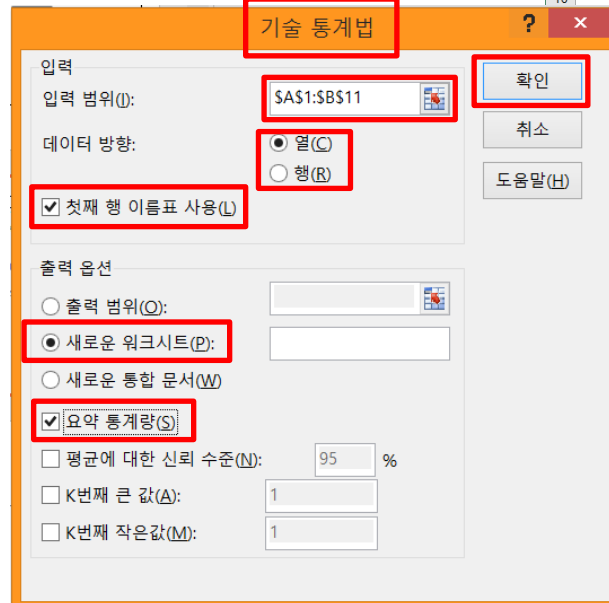


# KESS 활용한 분석 결과



# 데이터 분석을 활용한 분석 방법

[데이터]탭→[분석]그룹 → [데이터 분석] →  
[기술 통계법]선택 → [확인] →  
[기술 통계법]옵션 창에서 [입력 범위]에 데  
이터 범위 지정 →[데이터 방향]지정 →  
[첫째 행 이름표 사용]체크 →  
[출력 옵션]을 [새로운 워크시트]로  
지정 →[요약 통계량]로  
체크 →[확인]선택



# KESS 활용한 분석 결과

처리		대조	
평균	19.7	평균	16.3
표준 오차	1.183685	표준 오차	0.760847
중앙값	20	중앙값	16.5
최빈값	22	최빈값	19
표준 편차	3.743142	표준 편차	2.406011
분산	14.01111	분산	5.788889
첨도	0.858193	첨도	-0.51899
왜도	-0.80305	왜도	-0.67968
범위	13	범위	7
최소값	12	최소값	12
최대값	25	최대값	19
합	197	합	163
관측수	10	관측수	10

- KESS랑 데이터 분석을 활용한 경우의 차이점은 KESS에서는 출력하고 싶은 통계량만 선택할 수 있지만 데이터 분석의 경우는 특정 통계량만을 출력할 수는 없음

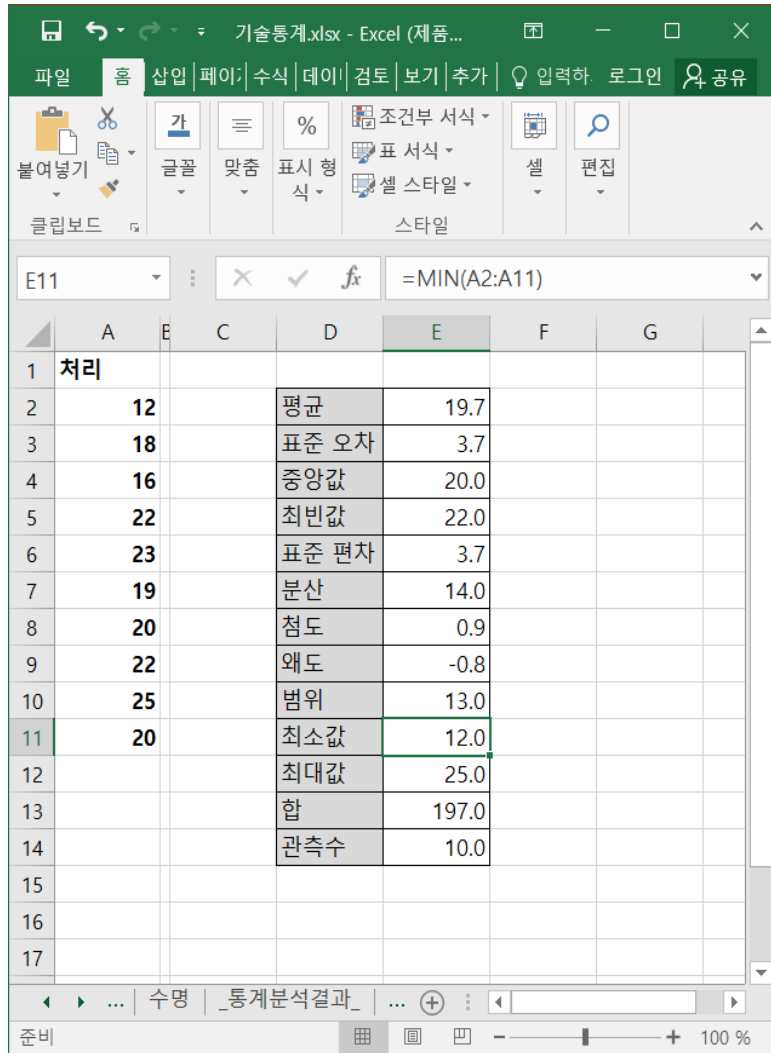
## <KESS>

The KESS interface displays various statistical analysis options. On the left, there are checkboxes for '히스토그램' (checked), '상자그림' (checked), '줄기잎그림' (unchecked), and '정규확률그림' (unchecked). Below these is a '모두선택' button. In the center, there are two sections: '기초통계량' (Basic Statistics) and '추가통계량' (Additional Statistics). The '기초통계량' section includes checkboxes for '평균' (checked), '합계' (unchecked), '개수' (unchecked), '중앙값' (checked), '표준편차' (checked), and '평균의 표준오차' (checked). The '추가통계량' section includes checkboxes for '최빈값' (checked), '분산' (checked), '사분위수 범위' (checked), '최소값' (checked), '변동계수' (checked), '왜도' (checked), and '첨도' (checked). A '선택취소' button is located at the bottom right of the '추가통계량' section.

## <데이터 분석>

The '데이터 분석' interface shows output options. At the top, there is a '출력 옵션' (Output Options) section. It includes radio buttons for '출력 범위(O):' (unchecked), '새로운 워크시트(P):' (checked), and '새로운 통합 문서(W):' (unchecked). Below these are checkboxes for '요약 통계량(S):' (checked), '평균에 대한 신뢰 수준(N):' (unchecked, with a value of 95 %), 'K번째 큰 값(A):' (unchecked, with a value of 1), and 'K번째 작은 값(M):' (unchecked, with a value of 1). There is a '선택취소' button at the bottom right.

# 함수를 활용한 분석 방법 및 결과



	A	C	D	E	F	G
1	처리					
2	12		평균	19.7		
3	18		표준 오차	3.7		
4	16		중앙값	20.0		
5	22		최빈값	22.0		
6	23		표준 편차	3.7		
7	19		분산	14.0		
8	20		첨도	0.9		
9	22		왜도	-0.8		
10	25		범위	13.0		
11	20		최소값	12.0		
12			최대값	25.0		
13			합	197.0		
14			관측수	10.0		

- 평균 = AVERAGE(A2:A11)
- 표준 오차 = STDEV(A2:A11)/  
SQRT(COUNT(A2:A11))
- 중앙값 = MEDIAN(A2:A11)
- 최빈값 = MODE(A2:A11)
- 표준 편차 = STDEV(A2:A11)
- 분산 = VAR(A2:A11)
- 첨도 = KURT(A2:A11)
- 왜도 = SKEW(A2:A11)
- 범위 = MAX(A2:A11) - MIN(A2:A11)
- 최소값 = MIN(A2:A11)
- 최대값 = MAX(A2:A11)
- 합 = SUM(A2:A11)
- 관측수 = COUNT(A2:A11)

# mpg 데이터

## <데이터 설명>

- 미국의 자동차 연비 측정 데이터(1999~2008)
- R프로그램의 ggplot2 패키지 제공 데이터로써 11개의 변수와 234개 차종의 관측값
  - trans(변속기 유형), model(차종), year(연식), class(차 분류) 등이 기록되어 있음

## <목적>

변속기 유형(trans)을 크게 2가지로 분류를 한 후 각각에 대한 기술통계를 분석하고자 함

## <분석 과정>

- ① 변속기 유형(trans)의 값 확인
- ② 변속기 유형(trans)의 값을 자동(manual), 수동(auto)으로만 구분하는 변수(re\_trans)생성
- ③ re\_trans에 대한 빈도 분석
- ④ re\_trans에 대한 기술통계 분석(상자그림 포함)

## 2.2 교차 분석

### 정의

- ✓ 두 개의 범주형 변수에 대한 빈도를 통하여 변수간의 관련성을 통계적으로 검정하는 방법

### 목적

- ✓ 교차 분석시 알아야하는 용어를 이해하고 교차 분석을 하는 방법을 파악하기 위함

### Focus!

- ✓ 교차 분석을 이해하기 위해 알아야 하는 용어는 무엇인가?
- ✓ 교차 분석은 무엇인가?

# Focus1 교차분석을 이해하기 위해 알아야 하는 용어는 무엇인가?

열 변수	1	2	· · ·	C	행 총합
행 변수	1	2	· · ·	C	행 총합
1	$n_{11}$	$n_{12}$	· · ·	$n_{1c}$	$n_{1.}$
2	$n_{21}$	$n_{22}$	· · ·	$n_{2c}$	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
R	$n_{r1}$	$n_{r2}$	· · ·	$n_{rc}$	$n_{r.}$
열 총합	$n_{.1}$	$n_{.2}$	· · ·	$n_{.c}$	n

$$n_{i.} = \sum_j n_{ij}$$

$$n_{.j} = \sum_i n_{ij}$$

$$n = \sum_i \sum_j n_{ij}$$

- $n_{11}$  : 1행, 1열의 값으로, 행 변수가 1이고 열 변수가 1인 경우에 대한 빈도
- 기대빈도 : 행 변수와 열 변수가 독립이라는 가정하에 R행, J열의 예상 빈도로  $E_{rc}$ 라 표시하고  $\frac{n_{.c} \times n_{r.}}{n}$ 로 계산
- 관측빈도 : 표본으로부터 관측된 빈도로  $O_{rc}$ 라 표시

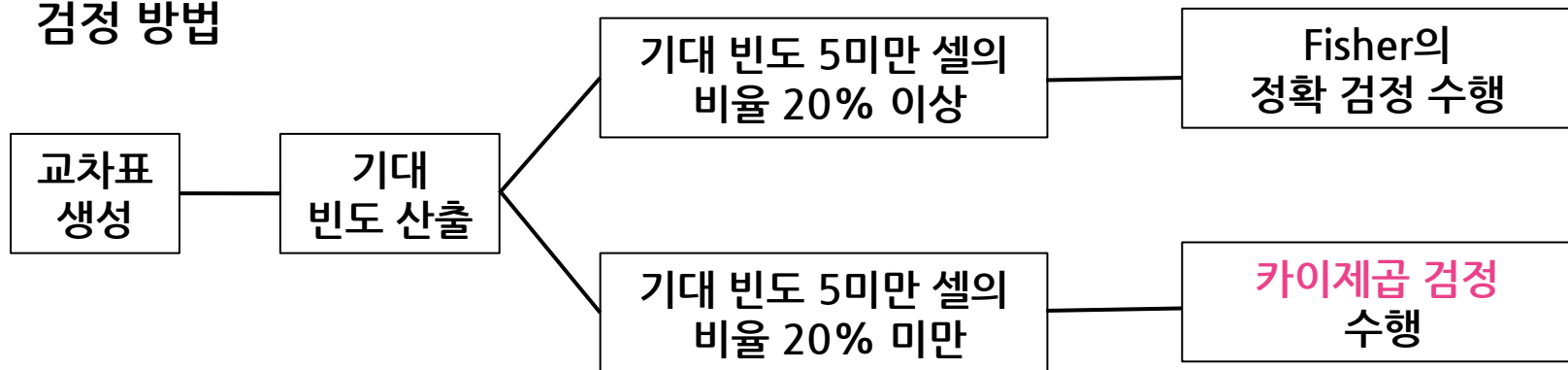


# Focus1 교차분석은 무엇인가?

## 교차 분석

- 정의 두 개의 범주형 변수에 대하여 독립성 여부 판단
- 가설 귀무가설( $H_0$ ) : 행 변수와 열 변수는 관련이 없다.  
대립가설( $H_1$ ) : 행 변수와 열 변수는 관련이 있다.

## 검정 방법



# KESS를 활용한 분석 방법

## <데이터 주의사항>

	A	B	C	D
1	구분	만족	보통	불만
2	남자	2	2	3
3	여자	4	3	4

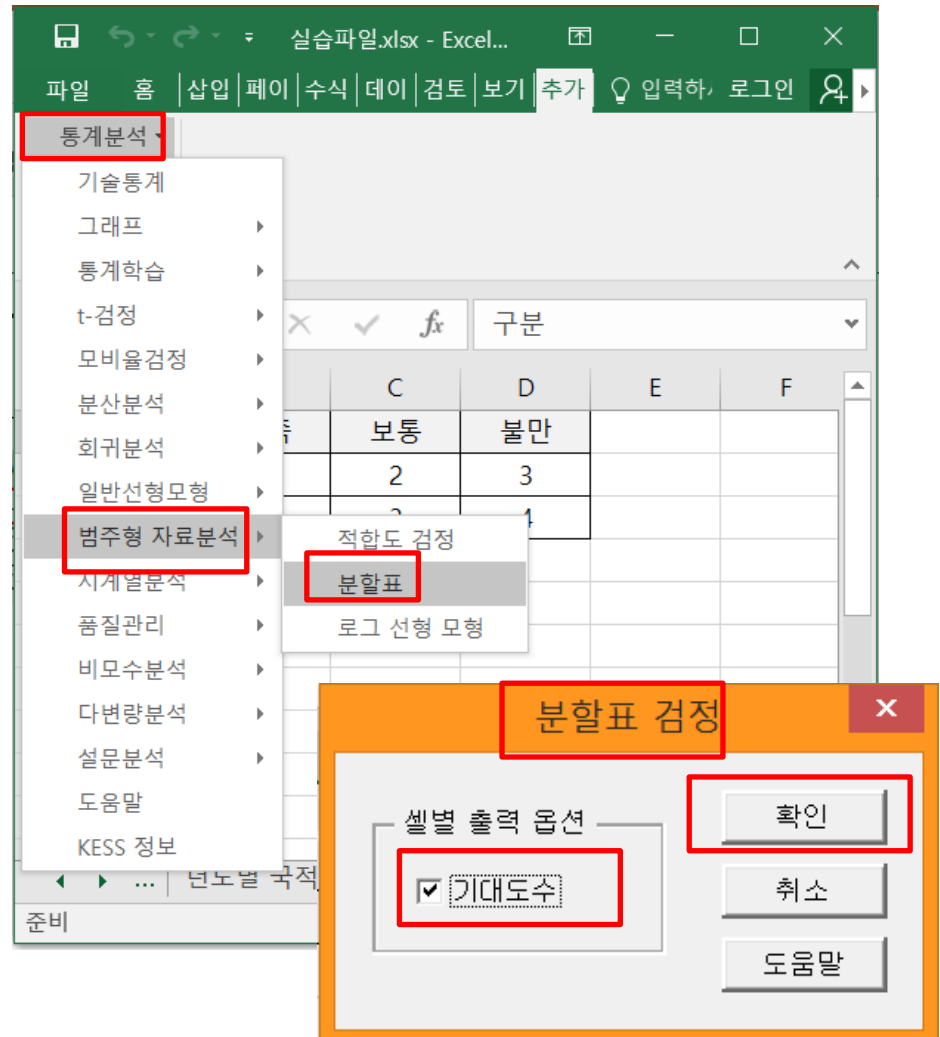
- A1셀부터 시작되어야 함
- 공백이 없어야 함(0으로 값 입력 필요)

[추가 기능] 탭 → [메뉴 명령] 그룹 →

[통계분석] → [범주형 자료분석] →

[분할표] → [분할표 검정] 옵션 창에서

[기대도수] 체크 → 확인



# KESS 활용한 분석 결과

## 분할표 검정

### 분할표

	만족	보통	불만	계
남자				
관측도수	2	2	3	7
기대도수	2,3333	1,9444	2,7222	
여자				
관측도수	4	3	4	11
기대도수	3,6667	3,0556	4,2778	
계	6	5	7	18

카이제곱 통계량 : 0,1269

유의확률 : 0,93852

100,0000%의 셀의 기대도수가 5보다 작습니다.

# mtcars 데이터

## <데이터 설명>

- 1974년
- R프로그램의 car패키지 제공 데이터로써 11개의 변수와 32개 차종의 관측값  
- am(변속기 유형), cyl(실린더 수), gear(전진기어 수), hp(마력) 등이 기록되어 있음

## <목적>

변속기 유형(am)과 마력(hp)에 대한 독립성 검정  
(단, 마력은 150을 기준으로 이하, 초과로 구분)

## <분석과정>

- ① 마력(hp)의 값을 이하( $hp \leq 150$ ), 초과( $hp > 150$ )로 구분하는 변수(re\_hp)생성
- ② 변속기 유형(am)과 마력(re\_hp)에 대한 교차표 생성
- ③ 변속기 유형(am)과 마력(re\_hp)에 대한 교차 분석 수행

## 2.3 평균 차이 검정

**정의**     ✓    하나 또는 두 개의 집단에 대한 평균 차이를 통계적으로 검정하는 방법

**목적**        ✓    평균 차이 검정이 어떻게 분류가 되어지고, 그 중에서도 단일/대응/독립 평균 차이 검정 방법을 파악하기 위함

**Focus!**    ✓    평균 차이 검정은 어떻게 분류되는가?  
              ✓    단일 표본 평균 차이 검정은 무엇인가?  
              ✓    대응 표본 평균 차이 검정은 무엇인가?  
              ✓    독립 표본 평균 차이 검정은 무엇인가?

# Focus1 평균 차이 검정은 어떻게 분류되는가?

## ① 정규성 검정

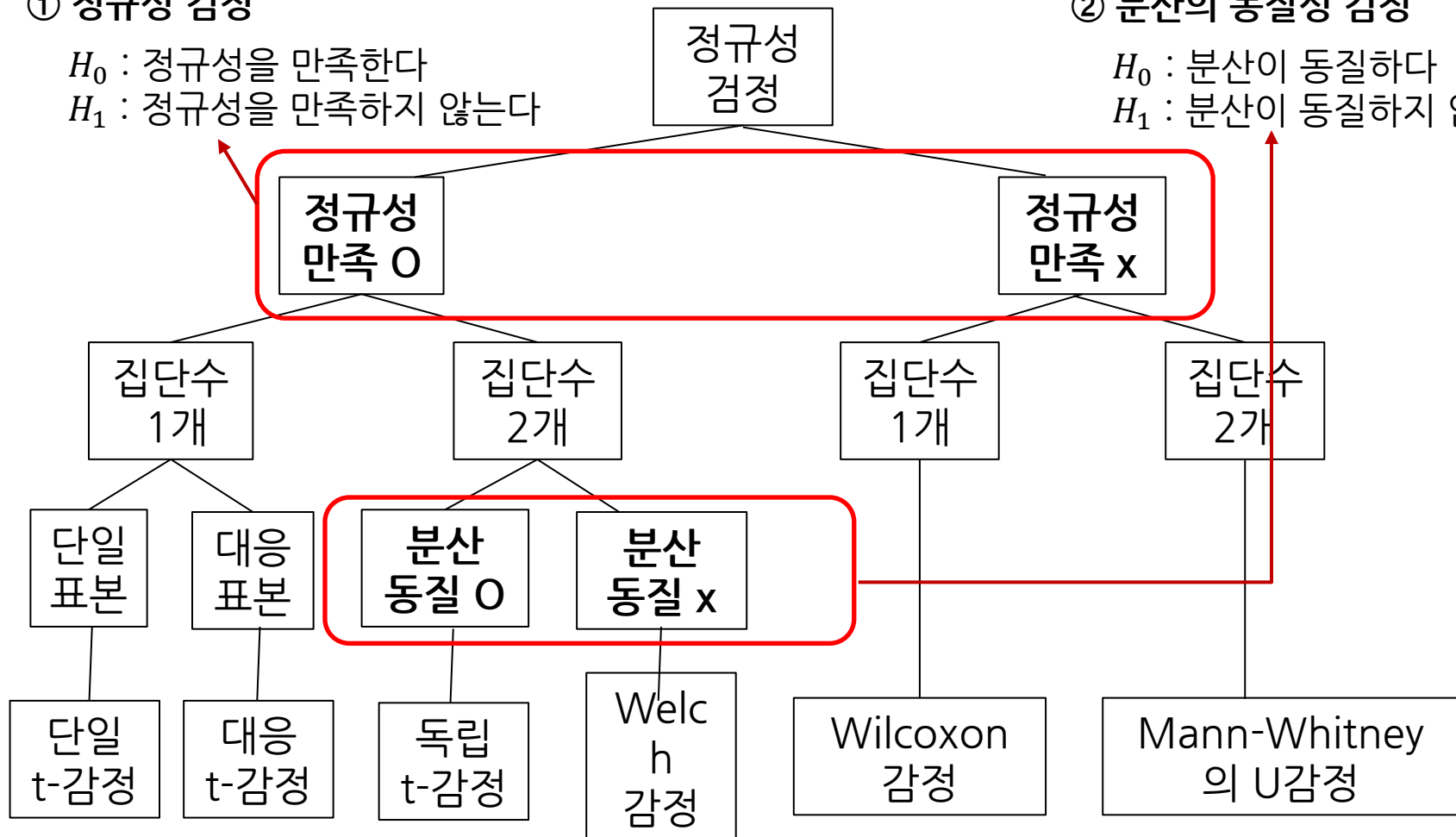
$H_0$  : 정규성을 만족한다

$H_1$  : 정규성을 만족하지 않는다

## ② 분산의 동질성 검정

$H_0$  : 분산이 동질하다

$H_1$  : 분산이 동질하지 않다



## Focus2 단일 표본 평균 차이 검정은 무엇인가?

🐼: 키가 180cm이니?

🐼: 고등학교 이후로 평균 182cm 유지 중이야, **우리반 애들 평균 키가 178cm인데!!**

🐼: 학기 초에 비해 애들이 다 많이 키가 컸구나?

🐼: 아니야, 옆 반 애들은 평균 키가 173cm일 걸? 대체적으로 우리 반 애들보다 작아

🐼: 아 그래? 너네 반 애들이 옆 반보다 평균적으로 키가 크다는 거지?

- 남자의 **반 평균 키가 178cm가 맞는지** 검정이 필요한 순간!

분석 집단은 남자의 반이며, 남자의 반의 평균 키가 178cm(특정 값)인지 아닌지 검정을 해야함

즉, **한 집단에 대한 평균과 특정 평균 값의 비교**가 필요

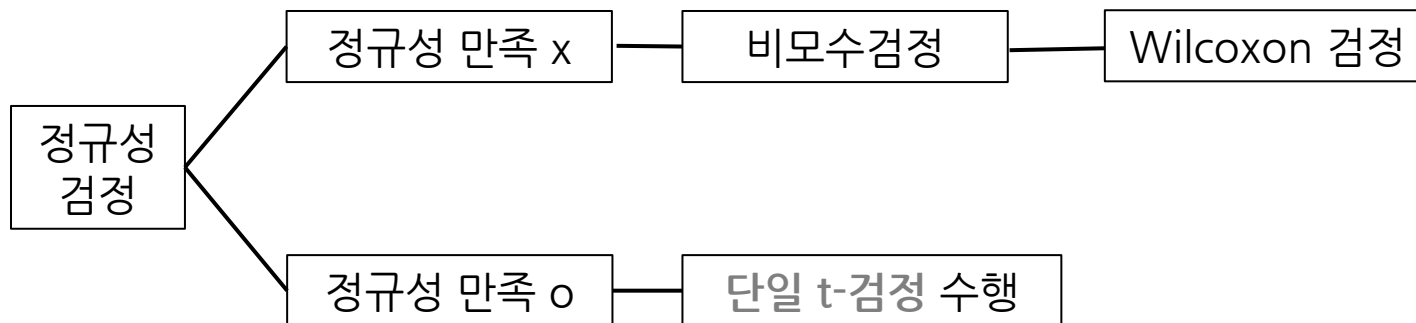
➡ **단일 표본 평균 차이 검정**

# Focus2 단일 표본 평균 차이 검정은 무엇인가?

## 단일 표본 평균 차이 검정

- 정의 한 집단에 대한 평균과 특정 평균 값의 비교
- 가설 귀무가설( $H_0$ ) : 남자의 반 평균 키는 178cm이다.  
대립가설( $H_1$ ) : 남자의 반 평균 키는 178cm가 아니다.

### 검정 방법 구분





## Focus3 대응 표본 평균 차이 검정은 무엇인가?

🐼: 키가 180cm이니?

🐼: 고등학교 이후로 평균 182cm 유지 중이야, 우리반 애들 평균 키가 178cm인데!!

🐼: 학기 초에 비해 애들이 다 많이 키가 컸구나?

🐼: 아니야, 옆 반 애들은 평균 키가 173cm일 걸? 대체적으로 우리 반 애들보다 작아

🐼: 아 그래? 너네 반 애들이 옆 반보다 평균적으로 키가 크다는 거지?

- 남자의 반 친구들이 학기 초에 비해 현재의 키가 큰지 검정이 필요한 순간!

분석 집단은 남자의 반이며, 학기 초 평균 키보다 현재 평균 키가 큰지 작거나 같은지 검정을 해야함

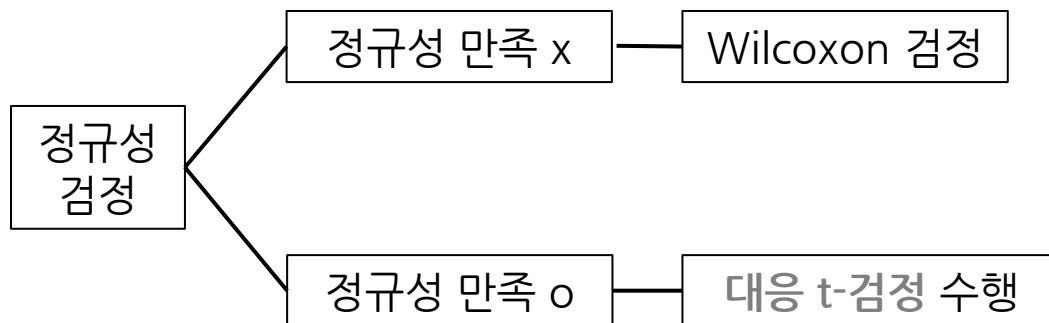
즉, 하나의 집단에 대한 두가지 시점의 평균 비교가 필요

➡ 대응 표본 평균 차이 검정

## Focus3 대응 표본 평균 차이 검정은 무엇인가?

### 대응 표본 평균 차이 검정

- 정의 하나의 집단에 대한 두가지 시점의 평균 비교
- 가설 귀무가설( $H_0$ ) : 남자의 반의 학기초 평균 키와 현재 평균 키는 작거나 같다  
대립가설( $H_1$ ) : 남자의 반의 학기초 평균 키와 현재 평균 키가 크다
- 검정 방법 구분



## Focus4 독립 표본 평균 차이 검정은 무엇인가?

🐼: 키가 180cm이니?

🐼: 고등학교 이후로 평균 182cm 유지 중이야, 우리반 애들 평균 키가 178cm인데!!

🐼: 학기 초에 비해 애들이 다 많이 키가 컸구나?

🐼: 아니야, 옆 반 애들은 평균 키가 173cm일 걸? 대체적으로 우리 반 애들보다 작아

🐼: 아 그래? 너네 반 애들이 옆 반보다 평균적으로 키가 크다는 거지?

- 남자의 반 평균 키가 옆 반의 평균 키보다 큰지 검정이 필요한 순간!

분석 집단은 남자의 반과 옆 반이며, 남자의 반이 옆 반보다 평균키가 큰지 작거나 같은지 검정을 해야함

즉, 두 개의 집단에 대한 평균 비교가 필요

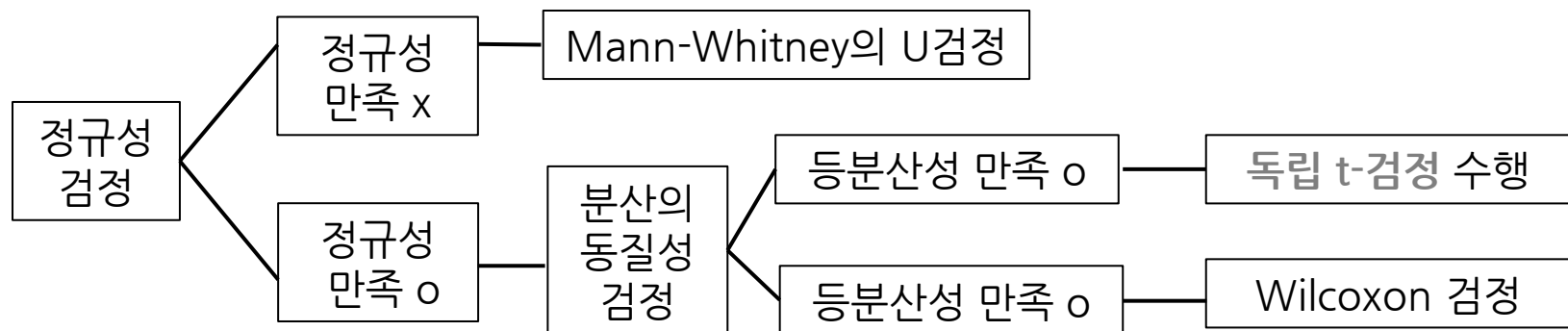
➡ 독립 표본 평균 차이 검정

# Focus4 독립 표본 평균 차이 검정은 무엇인가?

## 독립 표본 평균 차이 검정

- 정의 두 개의 집단에 대한 평균 비교
- 가설 귀무가설( $H_0$ ) : 남자의 반과 옆 반의 평균 키는 작거나 같다.  
대립가설( $H_1$ ) : 남자의 반과 옆 반의 평균 키가 크다.

### 검정 방법 구분



# KESS를 활용한 분석 방법

[추가 기능] 탭 → [메뉴 명령] 그룹 →

[통계분석] → [t-검정] → [일표본t-검정]

옵션 창에서 [분석변수]에 평균을 계산할

변수 추가 → [검정값]에 비교할 평균 수

치 입력 → [신뢰구간] 체크를 하면 값 변

경 가능 (기본값 : 95%) →

[대립가설] 지정 → [확인]선택



# KESS를 활용한 분석 결과

## t-검정 분석결과

### 일표본 검정

변수명	개수	평균	표준편차
score	30	170.3333	6.2275

$H_0 : \mu = \mu_0$  vs.  $H_1 : \mu < \mu_0$  ( $\mu_0 = 173$ )

t-통계량	자유도	유의확률
-2.3454	29	0.013

95% 신뢰구간 하한	상한
168.008	172.6587

# KESS를 활용한 분석 방법

[추가 기능] 탭 → [메뉴 명령] 그룹 →  
[통계분석] → [t-검정] → [이표본t-검정]  
옵션 창에서 [고급입력] 탭 선택 → [분류  
변수]에 집단 변수 추가, [분석 변수]에 평  
균을 계산할 변수 추가 → [대립가설] 지정  
→ [신뢰구간] 체크를 하면 값 변경 가능  
(기본값 : 95%) → [검정방법]에서 [독립  
비교] 선택  
→ [확인] 선택



# KESS를 활용한 분석 결과

## t-검정 분석결과

### 이표본 검정 (독립비교)

변수명	개수	평균	표준편차
A	30	72.5333	13.3073
B	30	76.4	12.1616

### 등분산 검정

자유도	F 값	유의확률
( 29 , 29 )	1.1973	0.631

"H0:두 표본의 분산들이 서로 같다."를 유의수준  $\alpha=0.05$ 에서 기각할 수 없다.  
 ※유의확률이 유의수준보다 큰 경우에는 등분산 결과를 사용하는 것이 좋다.

H :  $\mu_1 = \mu_2$  vs. K :  $\mu_1 \neq \mu_2$  ( $\mu_1$  : A,  $\mu_2$  : B )

분산	t-통계량	자유도	유의확률
등분산	-1.1748	58	0.2449
이분산	-1.1748	57.5361	0.2449



# KESS를 활용한 분석 방법

[추가 기능] 탭 → [메뉴 명령] 그룹 →  
[통계분석] → [t-검정] → [이표본t-검정]  
옵션 창에서 [고급입력] 탭 선택 → [분류  
변수]에 집단 변수 추가, [분석 변수]에 평  
균을 계산할 변수 추가 → [대립가설] 지정  
→ [신뢰구간] 체크를 하면 값 변경 가능  
(기본값 : 95%) → [검정방법]에서 [대응  
비교] 선택 → [확인] 선택



# KESS를 활용한 분석 결과

## t-검정 분석결과

### 이표본 검정 (대응비교)

변수명	개수	평균	표준편차
before	30	64.6	11.1033
after	30	61.5333	10.4971

$H_0 : \mu_1 - \mu_2 = 0$  vs.  $H_1 : \mu_1 - \mu_2 \neq 0$  ( $\mu_1$  : before,  $\mu_2$  : after )

t-통계량	자유도	유의확률
5.189	29	0

95% 신뢰구간 하한	상한
1.858	4.2754

## 2.4 분산 분석

**정의** ✓ 세 개 이상의 집단에 대한 평균 차이를 통계적으로 검정하는 방법

**목적** ✓ 분산 분석과 평균 차이 검정의 차이를 이해하고 분산 분석은 어떻게 분류되어지며 그 중에서 일원분류 분산분석은 무엇이고 여러 집단 중 특정 집단간의 유의한차이를 비교하는 다중비교가 무엇인지 파악하기 위함

**Focus!** ✓ 집단이 세 개 이상일 경우 왜 분산 분석으로 집단의 차이를 검정하는가?  
✓ 분산분석은 어떻게 분류되는가?  
✓ 인원분류 분산 분석은 무엇인가?  
✓ 다중비교는 무엇인가?

# Focus1 집단이 세 개 이상일 경우 왜 분산 분석으로 집단의 차이를 검정하는가?

❓ 평균 차이 검정을 사용하면 안되나?

평균 차이 검정으로 집단이 3개 이상일 경우를 검정할 수 없다  
평균 차이 검정을 여러 번 할 경우 그만큼 유의수준이 증가하기 때문

✓ 2개의 집단(A,B)에 대해 평균 차이 검정을 할 경우는 신뢰수준 95%이면,  
유의수준의  $1-0.95=5\%$  임

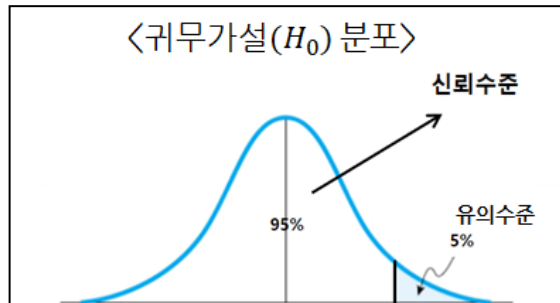
검정을 반복할 때, 신뢰수준은 귀무가설이 채택될 확률을 유의수준은 귀무가설이 기각될 확률을 의미함

✓ 3개의 집단(A,B,C)에 대해 평균 차이 검정을 할 경우는 평균 차이 검정을 3번 수행(A-B, A-C, B-C)하므로 신뢰수준이  $0.95*0.95*0.95=0.86$ 이 되고 **유의수준은  $1-0.86=14\%$ 가 됨**

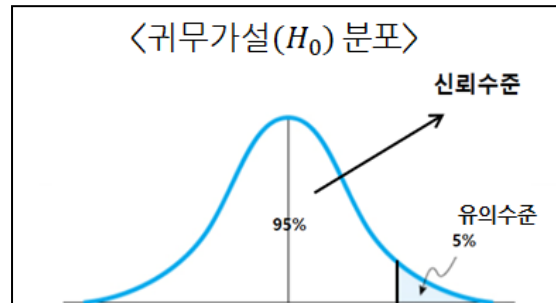
➡ 집단이 3개 이상일 경우, 평균 차이 검정을 2번 이상 수행하게 되는데,  
**평균 차이 검정 횟수가 증가함에 따라 신뢰 수준과 값이 변하기 때문에 부적절함**

# Focus1 집단이 세 개 이상일 겨우 왜 분산 분석으로 집단의 차이를 검정하는가?

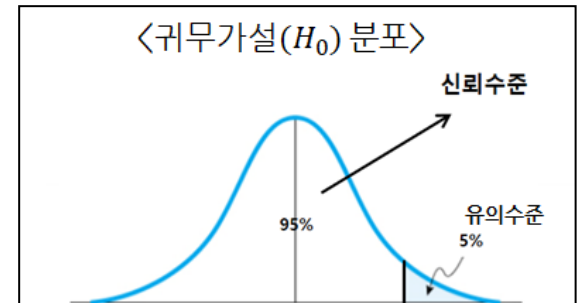
A집단 vs B집단



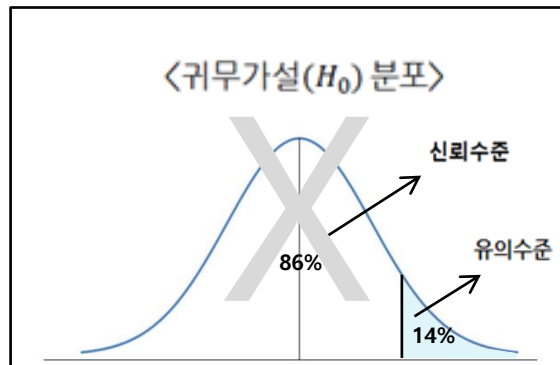
A집단 vs C집단



B집단 vs C집단

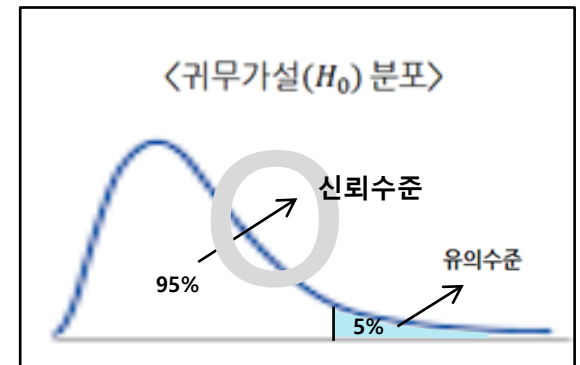


평균 차이 검정을 통한  
A집단 vs B집단 vs C집단








사전에 지정한 유의수준  
은 5%였는데 검정을 3번  
수행하면서 14%로 증가  
하게 됨

분산 분석을 통한  
A집단 vs B집단 vs C집단



## Focus3 일원분류 분산 분석은 무엇인가?

 : 무슨 소리야! 솔직히 다 비슷비슷하지~  
 : 말도안돼.. 어떻게 비슷비슷하다고 볼 수가 있지?  
 : 남자 반이 1반~3반이던가?  
 : 응. 우리 반이 1반, 재네 반이 2반 그리고 3반도 있지  
 : 그럼 1반~3반 평균 키를 비교해보자!!

- 남자의 1반~3반에 대한 평균 키가 차이가 있는지 검정이 필요한 순간!  
분석 집단은 남자의 (1반~3반)이며, 1반~3반의 평균 키가 동일한지 차이가 있는지 검정을 해야함  
즉, 두 개의 집단에 대한 평균 비교가 필요

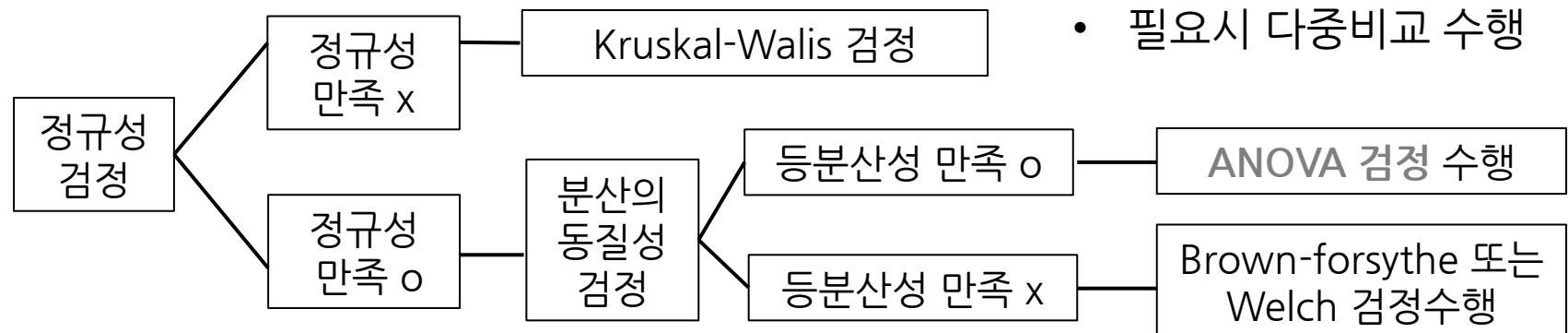
➡ 일원분류분산 분석

# Focus3 일원분류 분산 분석은 무엇인가?

## 일원분류 분산 분석

- 정의 세 개의 집단에 대한 평균 비교
- 가설 귀무가설( $H_0$ ) : 1~3반의 평균 키는 동일하다.  
대립가설( $H_1$ ) : 1~3반 중 적어도 어느 두 반 사이의 평균 키는 차이가 있다.

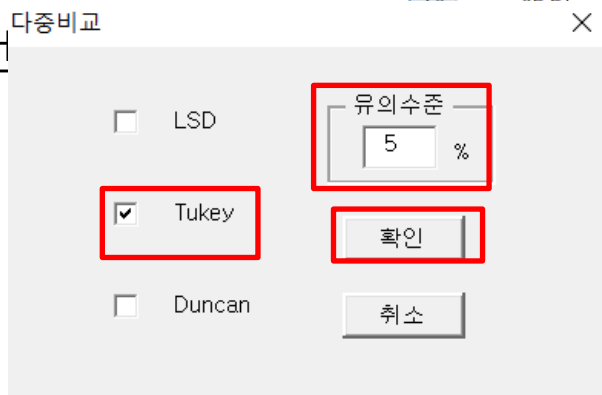
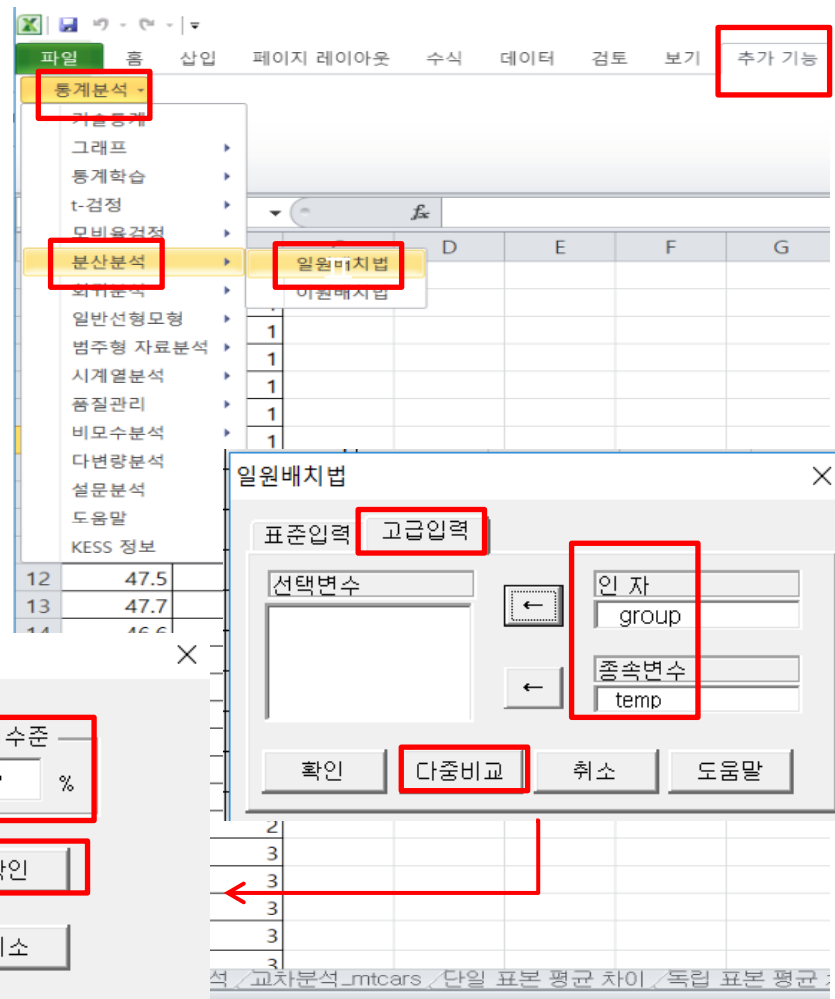
### 검정 방법 구분



- 필요시 다중비교 수행

# KESS를 활용한 분석 방법

[추가 기능] 탭 → [메뉴 명령] 그룹 →  
[통계분석] → [분산분석] → [일원배치  
법] 선택 → [일원배치법] 옵션 창에서 [고  
급입력] 탭 선택 → [인자]에 집단 변수 추  
가, [종속변수]에 평균을 계산할 변수 추  
가 → [다중비교] 선택 → [Tukey] 체크 →  
[유의수준] 설정 → [확인] 선택





# KESS를 활용한 분석 결과

## 일원배치 분산분석 결과

### 등분산 검정

Levene's test	제곱합	자유도	평균제곱	F값	유의확률
처리	0.3157	2	0.1579	0.7318	0.4903
잔차	5.8238	27	0.2157		

"H0:요인 수준 간 모분산들이 서로 같다."를 유의수준  $\alpha=0.05$ 에서 기각하지 못한다.  
즉, 모분산들이 ( $p>0.05$ ) 차이가 없다.  
등분산 가정이 만족 된다.

### 분산분석표

요인	제곱합	자유도	평균제곱	F값	유의확률
처리	156.302	2	78.151	108.8061	0
잔차	19.393	27	0.7183		
계	175.695	29			

"H0:모평균들이 서로 같다."를 유의수준  $\alpha=0.01$ 에서 기각한다.  
즉, 표본평균들이 아주 뚜렷한( $p<0.01$ ) 차이가 있다.

Tukey HSD		유의수준 = 0.05 에 대한 그룹		
group	자료수	그룹 1	그룹 2	
3	10	46.35		
2	10	46.94		
1	10		51.46	

같은 그룹에 속한 경우 유의수준  $\alpha=0.05$ 에서 처리평균에 차이가 없는 것으로 판단한다.

## 2.5 상관관계 분석

### 정의

- ✓ 두 개의 변수에 대해 한 변수가 증가함에 따라 다른 변수가 증가하거나 감소하는 **선형적인 상관관계**를 통계적으로 검정하는 방법

### 목적

- ✓ 상관관계 분석이 어떻게 분류가 되며, 상관관계 분석시 알아야하는 용어와 해석 방법을 이해하고 상관관계 분석 방법 중 피어슨의 상관계수를 파악하기 위함

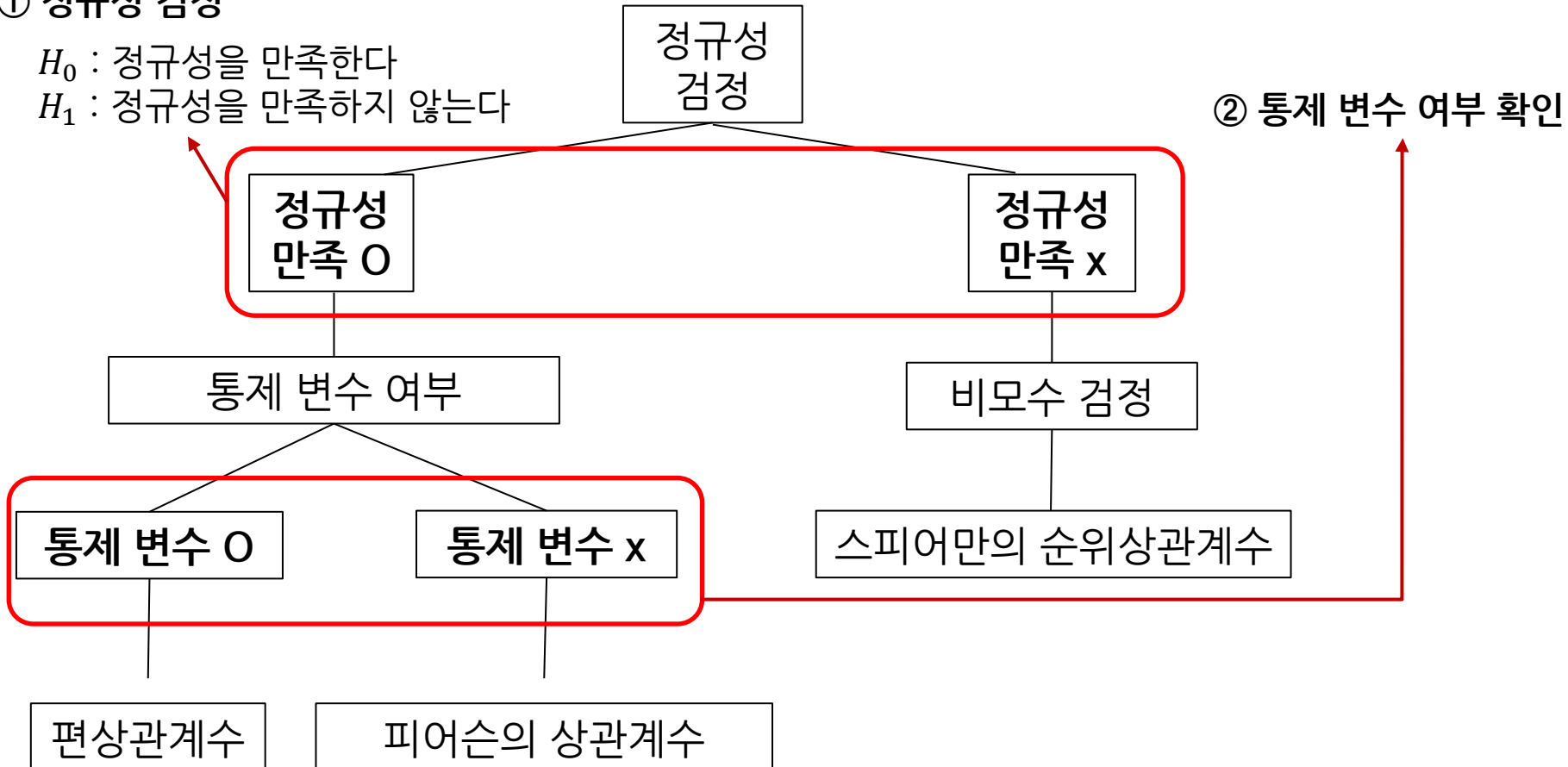
### Focus!

- ✓ 상관관계 분석은 어떻게 **분류**되는가?
- ✓ 상관관계 분석을 이해하기 위해 알아야 하는 **용어**는 무엇인가?
- ✓ 상관계수는 어떻게 **해석** 하는가?
- ✓ **피어슨의 상관계수**는 무엇인가?

# Focus1 상관관계 분석은 어떻게 분류 되는가?

## ① 정규성 검정

$H_0$  : 정규성을 만족한다  
 $H_1$  : 정규성을 만족하지 않는다



## ② 통제 변수 여부 확인

# Focus2 상관관계 분석을 이해하기 위해 알아야 하는 용어는 무엇인가?

## • 공분산

- ✓ 두 변수가 동시에 변하는 정도(상관관계)를 양으로 나타낸 척도로 두 변수 각각의 평균에 대해 얼마나 떨어져 있는지를 수치화한 값이며, 변수의 단위에 영향을 받음

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}, \quad 0 \leq s_{xy} \leq \infty$$

## • 상관계수

- ✓ 변수를 표준화하여 산출한 공분산과 동일하며 변수의 단위에 영향을 받지 않음 부호(+, -)는 방향(양, 음)을 나타내며, 크기는 상관관계의 강도를 나타냄

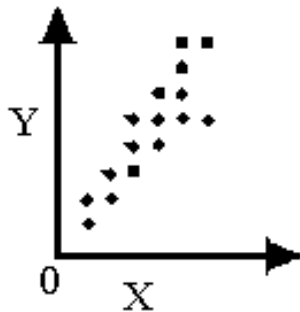
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad -1 \leq r_{xy} \leq 1$$

## Focus3 상관계수는 어떻게 해석하는가?

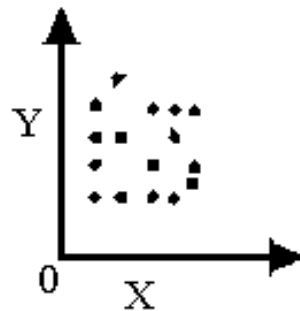
- 상관계수의 범위  $-1 < \text{상관계수} \leq 1$

$0 < \text{상관계수}(\rho) \leq 1$	한 변수가 증가함에 따라 다른 변수도 증가하는 경향을 나타냄
$\text{상관계수}(\rho) = 0$	두 변수간의 선형적인 관계를 가지고 있지 않음을 나타냄
$-1 \leq \text{상관계수}(\rho) < 0$	한 변수가 증가함에 따라 다른 변수는 감소하는 경향을 나타냄

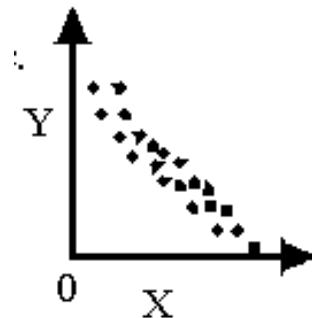
- 산점도



$0 < \text{상관계수}(\rho) \leq 1$



$\text{상관계수}(\rho) = 0$



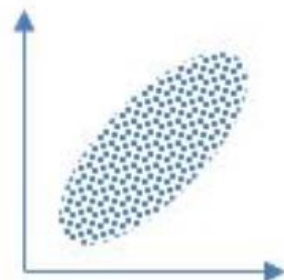
$-1 \leq \text{상관계수}(\rho) < 0$

## Focus3 상관계수는 어떻게 해석하는가?

- 상관계수에 대한 해석



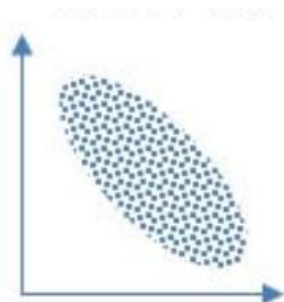
강한 양의 상관



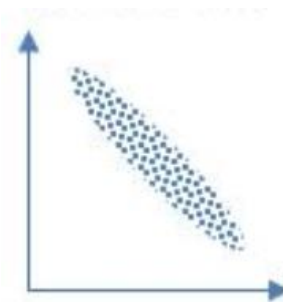
약한 양의 상관



상관 없음



약한 음의 상관

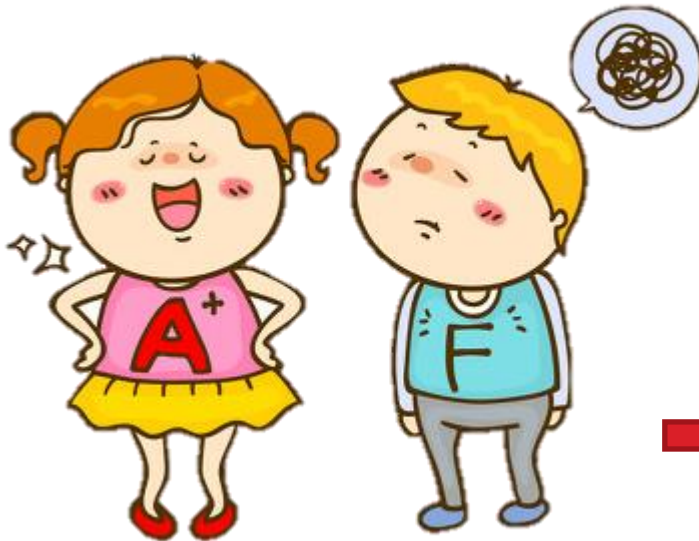


강한 음의 상관

## Focus3 상관계수는 어떻게 해석하는가?

① 상관계수가 인과관계를 나타낸다?

✓ 사례1. 성적과 자존감의 상관관계



• 성적과 자존감은 양의 상관관계가 있다

② 성적이 높아짐에 따라 자존감이 높아진다?  
③ 자존감이 높아짐에 따라 성적이 높아진다?



‘성적이 높아짐에 따라 자존감이 높아진다’,  
‘자존감이 높아짐에 따라 성적이 높아진다’  
모두 성립함 (상관관계O, 인과관계X)

## Focus3 상관계수는 어떻게 해석하는가?

① 상관계수가 인과관계를 나타낸다?

✓ 사례2. 키와 몸무게의 상관관계



• 몸무게와 키는 양의 상관관계가 있다

- ② 몸무게가 증가함에 따라 키가 커진다?
- ③ 키가 커짐에 따라 몸무게가 증가한다?



‘체중이 증가함에 따라 키가 커진다’,  
‘키가 커짐에 따라 몸무게가 증가한다’  
모두 성립함 (상관관계O, 인과관계X)



## Focus4 피어슨의 상관계수는 무엇인가?

① ? 혈압과 월급에는 상관관계가 있다?

- 혈압과 월급 사이의 상관관계 검정이 필요한 순간!

혈압과 월급이 수치형 변수일 때, 혈압과 월급에 대한 상관관계를 검정 해야함  
즉, 두 개의 수치형 변수에 대한 상관관계 분석이 필요

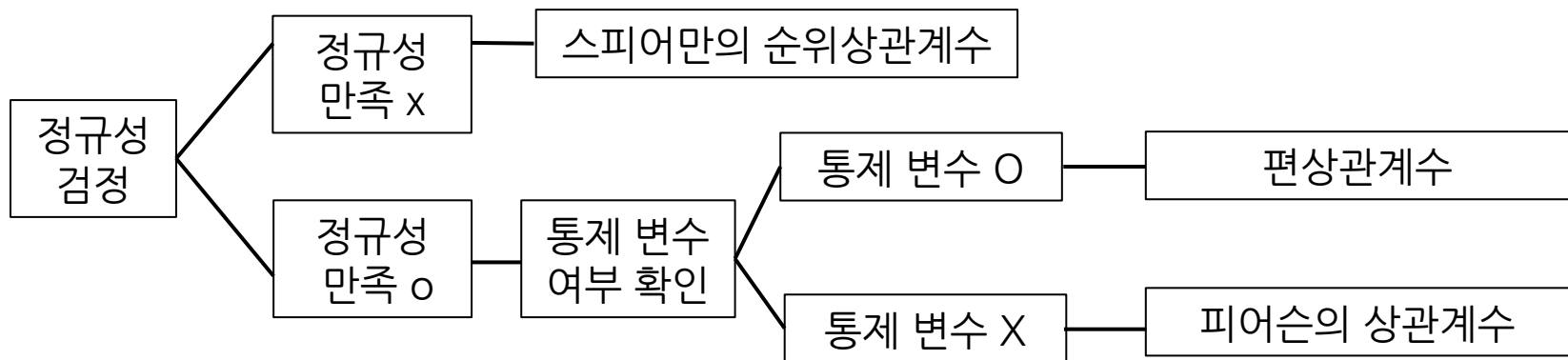
➡ 피어슨의 상관계수



# Focus4 피어슨의 상관계수는 무엇인가?

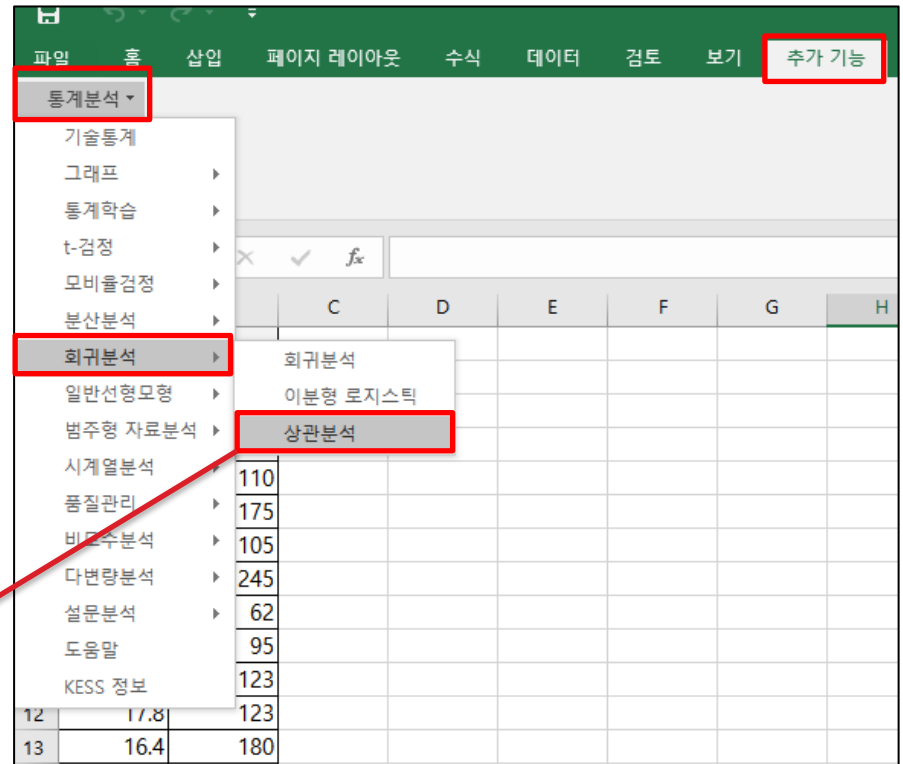
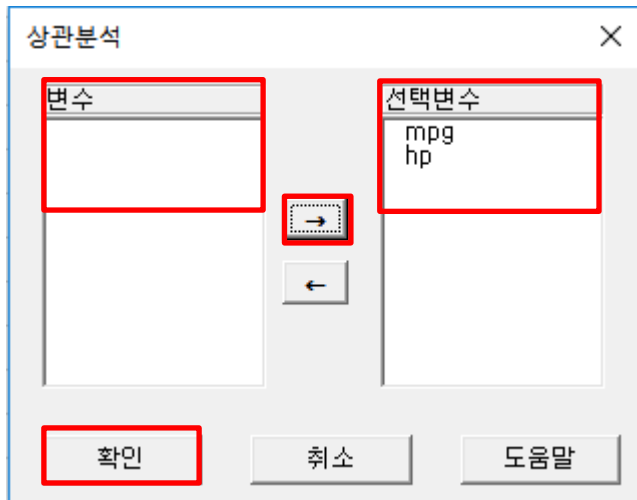
## 피어슨의 상관계수

- **정의** 두 개의 수치형 변수에 대해 한 변수가 증가함에 따라 다른 변수가 증가하거나 감소하는 선형적인 상관관계를 가지고 있는지 분석하는 방법
- **가설** 귀무가설( $H_0$ ) : 혈압과 월급간의 선형적 상관관계가 존재하지 않는다.  
대립가설( $H_1$ ) : 혈압과 월급간의 선형적 상관관계가 존재한다.
- **검정 방법 구분**



# KESS를 활용한 분석 방법

[추가 기능] 탭 → [메뉴 명령] 그룹 →  
[통계분석] → [회귀분석] → [상관분석] 선택  
→ [상관분석] 옵션창에서 [상관계수]에 상  
관계수를 계산할 변수 추가 → [확인] 선택



# KESS를 활용한 분석 결과

## 상관분석결과

### 상관분석

상관계수  
(유의확률)

	mpg	hp
mpg (유의확률)	1 .	-0,7762 0
hp (유의확률)	-0,7762 0	1 .

## 2.6 회귀 분석

### 정의

- ✓ 원인과 결과가 관계가 있는 변수에 대한 설명 변수가 증가함에 따라 반응 변수가 증가하거나 감소하는 선형적인 영향력을 통계적으로 검정하는 방법

### 목적

- ✓ 회귀 분석이 어떻게 분류가 되며, 모형은 어떻게 구성되어 있는지 데이터에 대한 기본 가정은 무엇인지를 이해하고 회귀 분석 방법 중 단순/다중 회귀 분석을 파악하기 위함

### Focus!

- ✓ 회귀 분석의 **종류**에는 어떤 것들이 있는가?
- ✓ **회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?**
- ✓ **단순 회귀 분석**은 무엇인가?
- ✓ **다중 회귀 분석**은 무엇이며 **변수 선택 방법**에는 어떤 것들이 있는가?
- ✓ 다중 회귀 분석에서 발생하는 **다중공선성**이란 무엇인가?

# Focus1 회귀 분석의 종류에는 어떤 것들이 있는가?

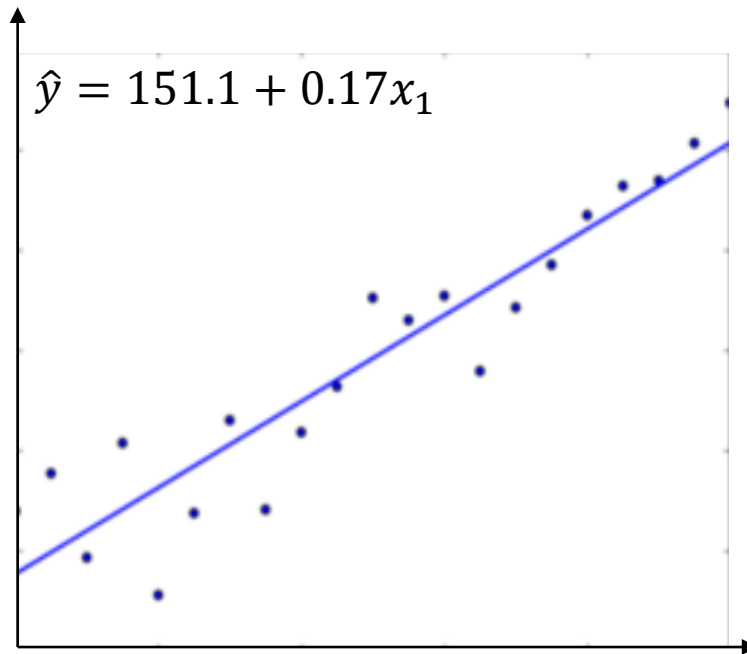
\* 반응 변수가 1개인 경우

반응 변수	설명 변수	회귀 분석
수치형 변수 1개	수치형/범주형 변수 1개	단순 회귀 분석
	수치형/범주형 변수 2개 이상	다중 회귀 분석
		라소 회귀 분석
		능형 회귀 분석
범주형 변수 1개	수치형/범주형 변수 1개 이상	로지스틱 회귀 분석

# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

## • 회귀 모형

$$Y = \beta_0 + \beta_1 x_{11} + \dots + \beta_p x_{pi} + \varepsilon_i$$



- $y$ (반응 변수) : 아들 키
- $x$ (설명 변수) : 아버지 키
- $i$ (개체 번호) : 25
- $p$ (설명 변수 개수) : 1
- $\beta_0$ (절편항) : 151.1
- $\beta_p$ (회귀계수) : 0.17
- $\varepsilon$ (오차항) : 설명할 수 없는  $y$ 의 변화량

# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

- 회귀 분석에서의 데이터 기본 가정

- ✓ 반응 변수와 설명 변수간 선형성

- ✓ 오차항에 대한 독립성

- ✓ 오차항에 대한 정규성

- ✓ 오차항에 대한 등분산성



# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

반응 변수와  
설명 변수간  
선형성

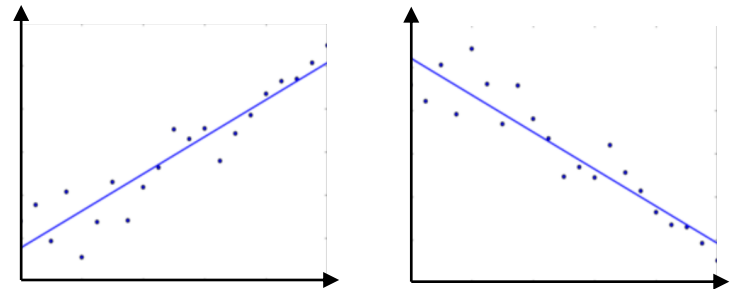
오차항에  
대한  
독립성

오차항에  
대한  
정규성

오차항에  
대한  
등분산성

- 반응 변수와 설명 변수는 선형성의 관계를 따라야 함

반응 변수와 설명 변수에 대한 산점도에서 관측값들이 회귀 직선을 중심으로 직선 형태를 보이면 선형성의 가정은 타당하다고 판단 가능



# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

반응 변수와  
설명 변수간  
선형성

오차항에  
대한  
독립성

오차항에  
대한  
정규성

오차항에  
대한  
등분산성

- 오차항은 독립적이어야 함

더빈-왓슨(DW) 통계량 값으로 판단 가능

$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

- 0에 가까우면 높은 양의 자기상관
- 2에 가까우면 자기상관이 존재하지 않음
- 4에 가까우면 높은 음의 자기상관

더빈-왓슨(DW) 통계량 값이 2에 가까우면 오차항에 대한 독립성 가정은 타당 하다고 판단 가능

# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

반응 변수와  
설명 변수간  
선형성

오차항에  
대한  
독립성

오차항에  
대한  
정규성

오차항에  
대한  
등분산성

- 오차항은 정규분포를 따라야 함

1. 오차로 그림 히스토그램에서 잔차들의 값이 정규분포 곡선에서 크게 벗어나지 않으면 오차항에 대한 정규성 가정이 충족되었다고 판단
2. 오차로 그린 정규확률 그림에서 잔차들의 값이 직선  $y=x$ 에서 크게 벗어나지 않으면 오차항에 대한 정규성 가정이 충족되었다고 판단

- 오차와 잔차는 집단의 차이
  - 모집단일 경우 : 오차
  - 표본 집단일 경우 : 잔차

# Focus2 회귀 모형은 어떻게 구성되어 있으며 기본 가정은 무엇인가?

반응 변수와  
설명 변수간  
선형성

오차항에  
대한  
독립성

오차항에  
대한  
정규성

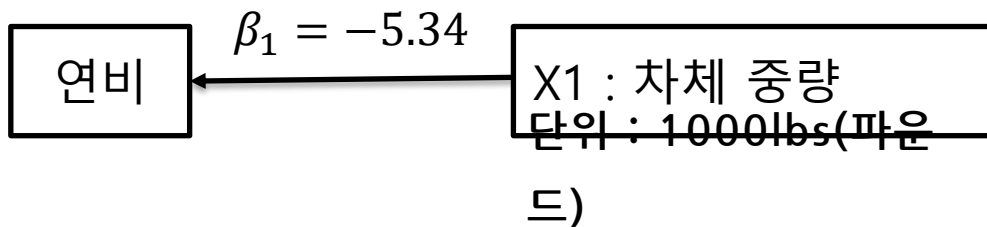
오차항에  
대한  
등분산성

- 오차항은 측정값에 동일한 분산이어야 함

예측된  $y$  값을  $x$  축으로 놓고 표준화 오차를  $y$  축으로 놓고 그린 산점도에서  $y$  축의 값이 0을 중심으로 랜덤하게 분포하면 오차항에 대한 등분산성 가정이 충족되었다고 판단

## Focus3 단순 회귀 분석은 무엇인가?

① 차체 중량이 연비에 영향을 미칠까?



차체 중량이 한 단위(1000파운드) 증가함에 따라 연비(갤런당 마일)가 5.34만큼 감소

- 차체 중량이 연비에 영향을 미치는지에 대한 검정이 필요할 때

수치형 반응 변수는 연비이고 설명 변수는 차체 중량으로 차체 중량이 연비에 영향을 미치는지 검정을 해야함

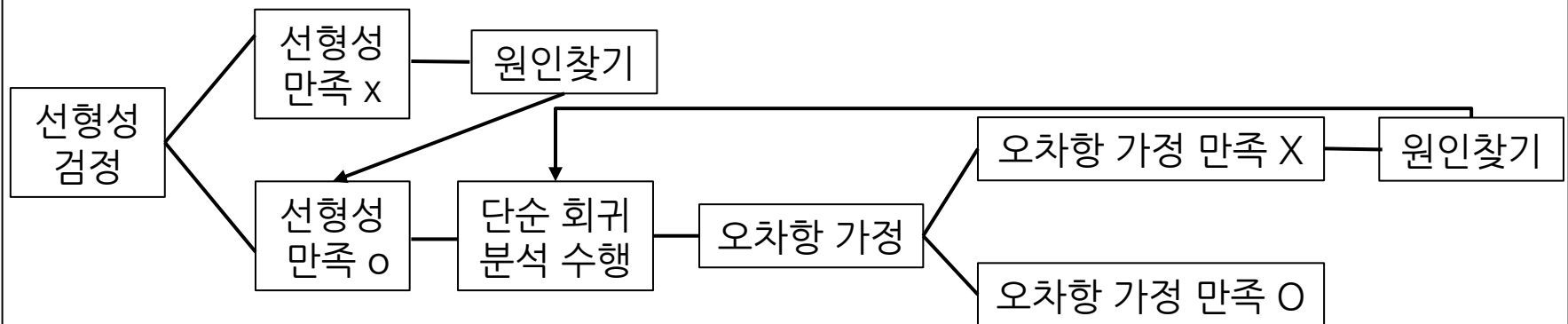
즉, 하나의 수치형 반응 변수와 하나의 설명 변수로 회귀 분석이 필요

➡ 단순 회귀 분석

# Focus3 단순 회귀 분석은 무엇인가?

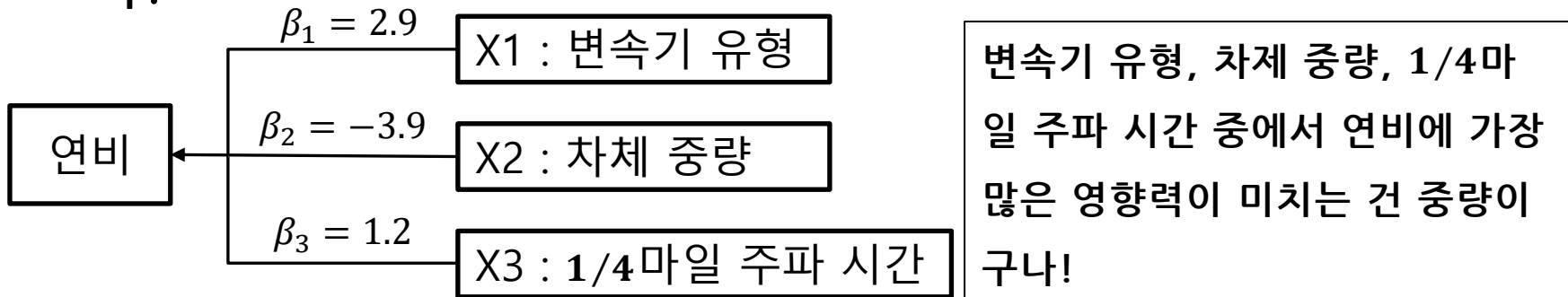
## 단순 회귀 분석

- **정의** 한 개의 설명 변수가 한 개의 수치형 반응 변수에 영향을 미치는지 분석하는 방법
- **가설** 귀무가설( $H_0$ ) : 차체 중량은 연비에 영향을 미치지 않는다.  
대립가설( $H_1$ ) : 차체 중량은 연비에 영향을 미친다.
- **분석 절차**



# Focus4 다중 회귀 분석은 무엇이며 변수 선택 방법에는 어떤 것들이 있는가?

① 변속기 유형, 차체 중량, 1/4마일 주파 시간이 연비에 영향을 미칠까?



- 변속기 유형, 차체 중량, 1/4마일 주파 시간이 연비에 영향을 미치는지에 대한 검정이 필요한 순간!

수치형 반응 변수는 연비이고 설명 변수는 변속기 유형, 중량, 1/4마일 주파 시간으로 3가지 설명 변수들이 연비에 영향을 미치는지 검정을 해야함  
즉, 하나의 수치형 반응 변수와 3개의 설명 변수로 회귀 분석이 필요

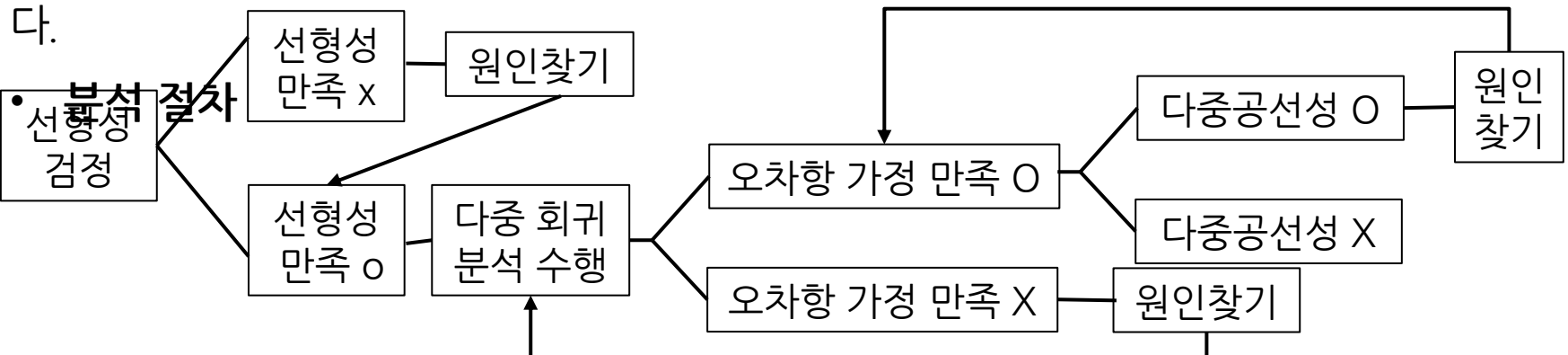
➡ 다중 회귀 분석

# Focus4 다중 회귀 분석은 무엇이며 변수 선택 방법에는 어떤 것들이 있는가?

## 다중 회귀 분석

- **정의** 두 개 이상의 설명 변수가 한 개의 수치형 반응 변수에 영향을 미치는지 분석하는 방법
- **가설** 귀무가설( $H_0$ ): 변속기 유형, 중량, 1/4마일 주파 시간은 연비에 영향을 미치지 않는다.

대립가설( $H_1$ ): 변속기 유형, 중량, 1/4마일 주파 시간 중 적어도 하나는 연비에 영향을 미친다.





# Focus4 다중 회귀 분석은 무엇이며 변수 선택 방법에는 어떤 것들이 있는가?

- 변수 선택 방법

하나의 반응 변수와 설명 변수들간의 조합이 다양할 때, 수정된 결정계수, 평균제곱 오차, AIC, BIC 등의 통계적인 기준을 통하여 적합한 설명 변수를 선택

- ☒ 전진적 선택 방법

- 고려하는 설명 변수들 중 가장 중요한 변수부터 차례로 모형에 추가하는 방법
- 문제점 : 모형에 추가된 후, 새로운 변수로 인해 중요도가 떨어져도 제거 안됨

- ☒ 후진적 제거 방법

- 모든 설명 변수를 포함한 후 가장 중요하지 않는 변수부터 차례로 제거하는 방법
- 문제점 : 한번 제거된 설명 변수는 다시 고려 안됨

- ☒ 단계적 선택 방법 (전진적 선택 방법 + 후진적 제거 방법)

- 가장 중요한 변수부터 추가시키며, 새로운 변수로 인해 필요 없게 되면 제거하는 방법
- 최적의 설명 변수들만 모을 수 있음

## Focus5 다중 회귀 분석에서 발생하는 다중공선성이란 무엇인가?

- 다중공선성

회귀 분석시 설명 변수들간의 유의한 상관관계가 존재하여 발생하는 문제를 의미하여 다중공선성을 측정하기 위한 변수의 상관관계를 조사하는 방법에는 분산확대인자(VIF), 상태지수가 있고, 다중공선성 문제를 해결하는 방법으로는 문제를 일으키는 설명 변수를 제거 또는 변환하거나 설명 변수를 축약하는 방법(주성분 분석, 요인 분석), 다중공선성을 고려하여 분석하는 회귀 분석 기법(라소 회귀 분석, 능형 회귀 분석)을 사용하는 방법이 있음

- ✓ 다중공선성 기준값

- 분산확대인자 : 일반적으로 10이상이면 설명 변수가 다중공선성 문제를 발생시킨다고 판단
- 상태지수 : 100이상인 값이 있으면 상관관계가 매우 유의한 설명 변수가 존재함을 의미하며 다중공선성 문제가 발생한다고 판단

The screenshot shows the Minitab '회귀분석(Regression)' dialog box. The '종속변수(Y)' is set to 'mpg' and the '독립변수(X)' is set to 'wt'. The '상수항 포함' checkbox is checked. The '변수 선택' button is highlighted.

# KESS를 활용한 분석 방법

회귀진단

잔차 | 표준화잔차 | 표준화제외잔차 |

☐ 데이터 시트에 출력

☒ vs 관측순서 그래프

☐ vs 예측치 그래프

☒ 히스토그램

☒ 정규확률그림

☐ 모두선택

진단통계량

☐ 다중공선성

☐ 영향관측점

☐ 부분회귀산점도

확인

취소

도움말

회귀진단

잔차 | 표준화잔차 | 표준화제외잔차 |

☐ 데이터 시트에 출력

☐ vs 관측순서 그래프

☒ vs 예측치 그래프

☐ 히스토그램

☐ 정규확률그림

☐ 모두선택

진단통계량

☒ 다중공선성

☐ 영향관측점

☐ 부분회귀산점도

확인

취소

도움말

변수선택

변수선택방법

☒ 변수 추가법

☐ 변수 제거법

유의수준

추가 5 %

변수선택

변수선택방법

☐ 변수 추가법

☒ 변수 제거법

☐ 변수 증감법

☐ 모든 가능한 회귀

유의수준

제거 10 %

변수선택

변수선택방법

☐ 변수 추가법

☐ 변수 제거법

☒ 변수 증감법

☐ 모든 가능한 회귀

유의수준

추가 15 %

제거 15 %

# KESS를 활용한 분석 결과

## 회귀분석결과

### 분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	847,7252	1	847,7252	91,375	< 0,0001
잔차	278,3219	30	9,2774		
계	1126,0472	31			

Root MSE      3,0459  
 결정계수      0,7528  
 수정결정계수    8,6622

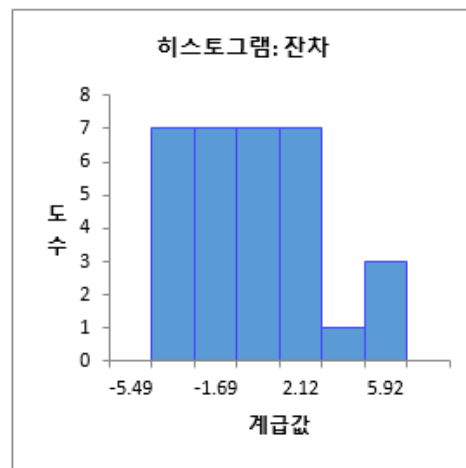
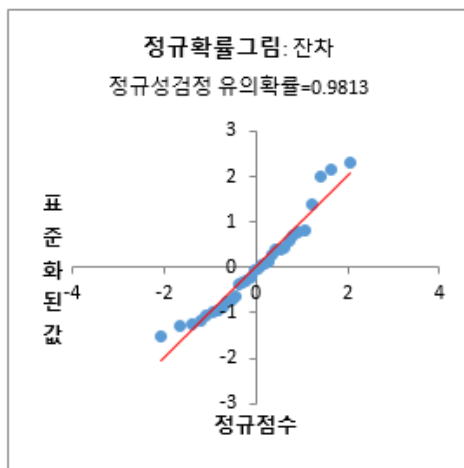
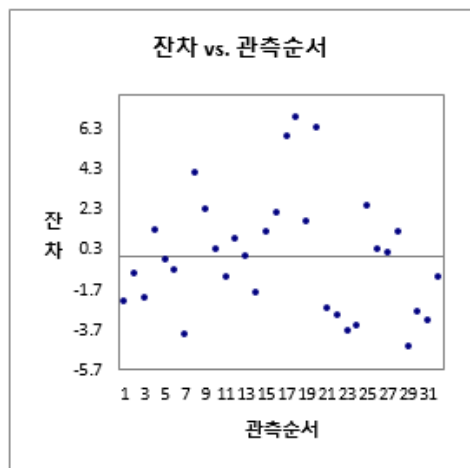
### 모수 추정

변수명	추정값	표준오차	t-통계량	유의확률
절편	37,28513	1,87763	19,858	< 0,0001
wt	-5,34447	0,55910	-9,559	< 0,0001

# KESS를 활용한 분석 결과

## 회귀진단결과

### 잔차그래프



DW 통계량 1차 자기상관계수

1.2517      0.3629

잔차들이 양의 자기상관을 가지면 DW통계량은 0에 가까운 값을 갖고  
음의 자기상관을 가지면 4에 가까운 값을 갖게 된다.

# KESS를 활용한 분석 결과

## 회귀진단결과

### 다중 공선성

변수명	분산팽창 인자
상수항	0,0000
cyl	15,3738
disp	21,6202
hp	9,8320
drat	3,3746
wt	15,1649
qsec	7,5280
vs	4,9659
am	4,6485
gear	5,3575
carb	7,9087

분산팽창인자 > 10 이면 다중공선성에 심각한 문제가 있다고 판정한다.

번호	고유값	상태지수	분산비율 상수항	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	9,0972	1,0000	0,0000	0,0001	0,0001	0,0002	0,0001	0,0001	0,0000	0,0005	0,0006	0,0001	0,0003
2	1,1284	2,8394	0,0000	0,0002	0,0012	0,0011	0,0001	0,0002	0,0000	0,0381	0,0282	0,0001	0,0010
3	0,5639	4,0164	0,0000	0,0000	0,0002	0,0004	0,0000	0,0003	0,0001	0,0614	0,1164	0,0001	0,0046
4	0,1158	8,8642	0,0001	0,0012	0,0076	0,0044	0,0007	0,0002	0,0003	0,1012	0,0603	0,0000	0,1639
5	0,0483	13,7225	0,0007	0,0002	0,0410	0,0996	0,0117	0,0002	0,0022	0,3358	0,1845	0,0080	0,0220
6	0,0220	20,3177	0,0004	0,0128	0,0399	0,2293	0,0095	0,0907	0,0001	0,0138	0,2612	0,0166	0,0725
7	0,0097	30,6824	0,0011	0,2450	0,0841	0,0135	0,0461	0,0001	0,0014	0,1059	0,2586	0,2077	0,0279
8	0,0063	37,9877	0,0005	0,0783	0,3625	0,5760	0,0674	0,1809	0,0260	0,2268	0,0238	0,0065	0,1746
9	0,0059	39,1545	0,0001	0,0346	0,0011	0,0627	0,5461	0,0001	0,0000	0,0108	0,0014	0,4823	0,0097
10	0,0020	68,0628	0,0255	0,2460	0,4534	0,0087	0,2028	0,6788	0,2220	0,0802	0,0045	0,1608	0,5233
11	0,0004	143,4209	0,9715	0,3816	0,0091	0,0040	0,1154	0,0484	0,7479	0,0254	0,0604	0,1176	0,0002

고유값이 평균 크기인 1에 비해 심각하게 작을 경우

상태지수의 값이 10보다 클 경우

분산비율이 80-90% 이상으로 나타나는 설명변수의 개수가 둘 이상인 경우

다중공선성의 문제가 있다고 판정한다.

# 감사합니다

---