2강. 데이터 분석과 통계

1. 기술통계와 추리통계

1.1 기술 통계 - 수집한 자료를 분석하여 대상들의 속성을 파악하는 통계 방법

- 중심경향값 : 전체 자료를 대표할 수 있는 수치들
- 분산도: 전체 자료가 얼마나 퍼져 있는 지를 알 수 있는 수치들
- 상관계수: 두 변수 간의 관계의 크기
- 회귀계수: 독립변수(원인)가 종속변수(결과)에 미치는 영향의 크기

1. 기술통계와 추리통계

1.1 기술 통계 - 수집한 자료를 분석하여 대상들의 속성을 파악하는 통계 방법

❖ 중심경향값

- 평균 : 전체 자료가 가지는 수치들의 총합을 전체 자료의 수로 나눈 수치
- 중앙값: 최대값과 최소값의 정가운데 수치
- 최빈값: 가장 많은 빈도를 보이는 수치

1. 기술통계와 추리통계

1.1 기술 통계 - 수집한 자료를 분석하여 대상들의 속성을 파악하는 통계 방법

❖ 중심경향값

- 평균 : 전체 자료가 가지는 수치들의 총합을 전체 자료의 수로 나눈 수치
- 중앙값: 최대값과 최소값의 정가운데 수치
- 최빈값: 가장 많은 빈도를 보이는 수치

1. 기술통계와 추리통계

1.1 기술 통계 - 수집한 자료를 분석하여 대상들의 속성을 파악하는 통계 방법

❖ 분산도

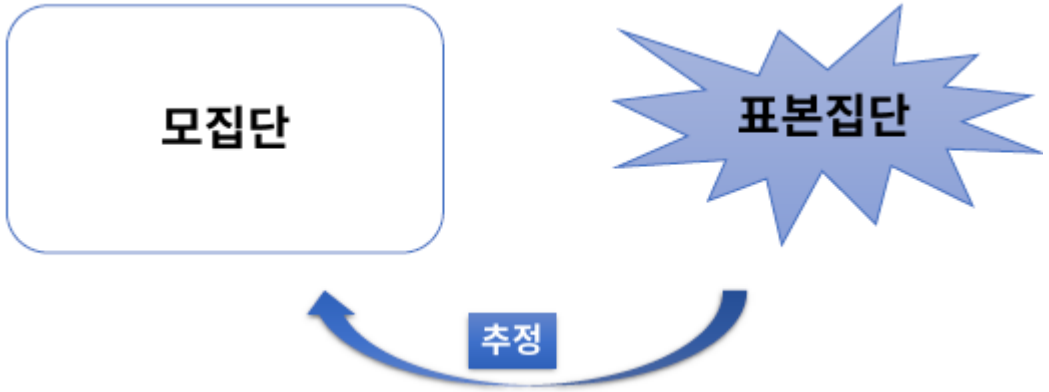
- 분산: 각 자료가 평균으로 부터 떨어진 거리(편차)들을 제곱한 수치들의 총합을 전체 자료의 수로 나눈 수치

- 표준편차: 분산을 제곱근을 취한 수치

1. 기술통계와 추리통계

1.2 추리 통계

- **모집단을 대표하는 표본을 추출하고** 표본의 기술통계를 이용하여 **모집단의 속성들을 유추**하는 통계방법



1. 기술통계와 추리통계

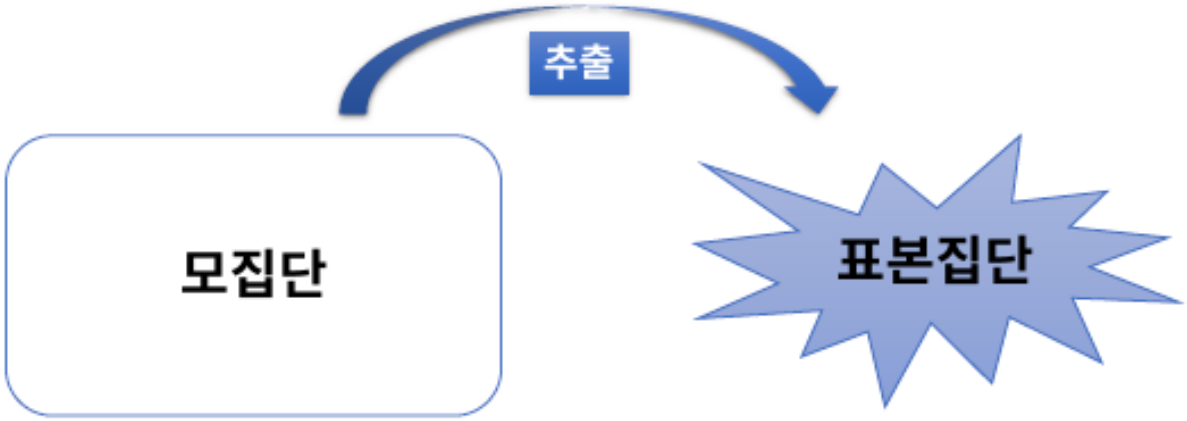
1.3 신뢰구간

- 추리통계에서 예측한 모집단의 특성이 위치할 가능성이 높은 구간
- 표본에서 구해지는 기술통계값들을 이용하여 계산되며, 95%, 99%, 99.9% 신뢰수준에서 따라 달라짐
- 95% 신뢰구간보다 99% 신뢰구간 영역이 더 넓음

2. 모집단과 표본

2.1 모집단

- 연구 또는 분석이 이루어지는 전체 대상



2. 모집단과 표본

2.2 표본 - 모집단에서 추출한 일부로 , 모집단의 속성들을 유추하는데 사용

- 확률표본추출 방법: 무작위로 표본을 추출하는 방법으로 모집단을 대표할 가능성이 높은 방법

- 비확률표본추출방법 : 조사자의 편의나 판단에 의해서 표본을 추출하는 방법으로 모집단을 대표하지 않을 가능성이 존재하는 방법

2. 모집단과 표본

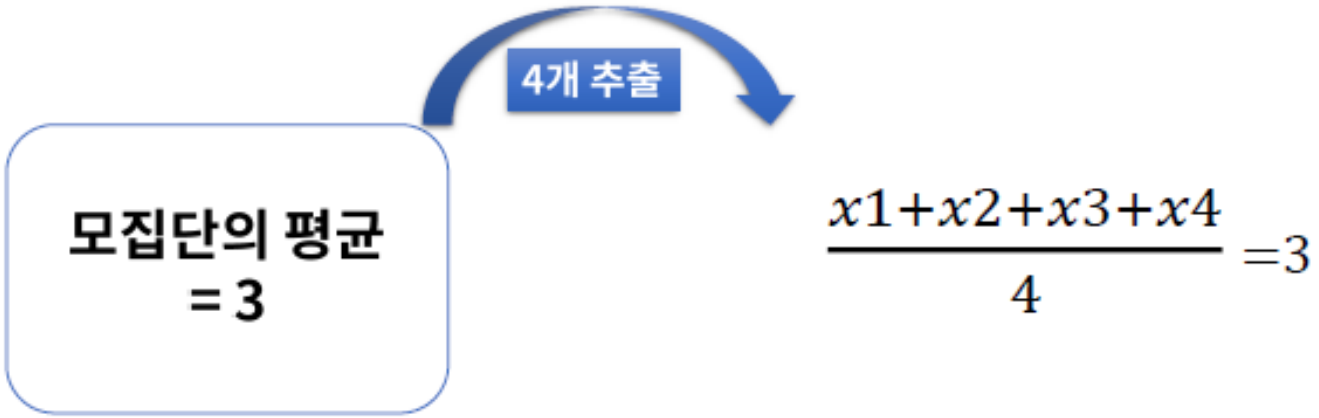
2.3 중심극한정리 - 표본이 30 이상으로 충분히 클 때

- 모집단의 분포와 상관없이 표본은 정규분포
- 표본의 평균 = 모집단의 평균
- 표본의 분산 = (모집단의 분산)/(표본의 수)

2. 모집단과 표본

2.4 자유도

- 평균을 유지하면서 자유롭게 어떠한 값도 가질 수 있는 사례의 수 (N-1)



2. 모집단과 표본

2.4 자유도

- 평균을 유지하면서 자유롭게 어떠한 값도 가질 수 있는 사례의 수 (N-1)

$$t-value=\frac{\left(\bar{Y}_A-\bar{Y}_B\right)-\left(\mu_A-\mu_B\right)}{\sigma_{\left(\bar{Y}_A-\bar{Y}_B\right)}}=\frac{\left(\bar{Y}_A-\bar{Y}_B\right)}{\sqrt{\frac{\left(N_A-1\right)s_A^2+\left(N_B-1\right)s_B^2}{\left(N_A-1\right)+\left(N_B-1\right)}}\cdot\sqrt{\frac{1}{N_A}+\frac{1}{N_B}}}$$

$$F-value=\frac{MS_B}{MS_W}=\frac{SS_B/\left(J-1\right)}{SS_W/\left(N-J\right)}=\frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{Y}_j-\bar{Y}\right)^2/\left(J-1\right)}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\bar{Y}_{ij}-\bar{Y}_j\right)^2/\left(N-J\right)}$$

3. 척도

3.1 척도의 원칙

- **포괄성: 척도 안에 모든 경우의 수가 포함되어야 한다는 원칙**

당신의 최종 학력은?

(1) 고졸 (2) 전문대졸 (3) 4년대졸 (4) 대학원졸

당신의 최종 학력은?

(1) 고졸 이하 (2) 전문대졸 (3) 4년대졸 (4) 대학원졸 이상

3. 척도

3.1 척도의 원칙

- **상호배타성: 척도 안에 중복되는 경우의 수가 없어야 한다는 원칙**

당신의 월급은?

(1) 1백만원 이하 (2) 1백만원 ~ 2백만원
(3) 2백만원 ~ 3백만원 (4) 3백만원 이상

당신의 월급은?

(1) 1백만원 이하 (2) 1백만원 초과 ~ 2백만원 이상
(3) 2백만원 초과 ~ 3백만원 이상 (4) 3백만원 이상

3. 척도

3.2 명목척도

- **측정이 이루어지는 항목들이 상호배타적인 특성만을 가진 척도**

당신의 성별은?

(1)남성 (2)여성

3. 척도

3.3 서열척도

❖ **명목척도들 중 항목들 간에 서열이나 순위가 존재하는 척도**

당신의 최종 학력은?

(1) 무학 (2) 초졸 (3) 중졸 (4)고졸 (5) 전문대졸

(6) 4년대졸 (7) 석사졸 (8) 박사졸 (9) 기타

3. 척도

3.4 등간척도

❖ **서열척도들 중 항목들 간의 간격이 일정한 척도**

당신의 직무에 대해 얼마나 만족하십니까?

(1) 전혀 만족하지 못한다 (2) 거의 만족하지 못한다

(3) 보통이다 (4) 약간 만족한다 (5) 매우 만족한다

3. 척도

3.5 비율척도

- **등간척도 중 아무 것도 없는 상태를 '0'으로 정할 수 있는 척도**

당신의 연령은 ?

(1)20세 이하 (2) 21 ~ 30 (3) 31 ~ 40 (4) 41 ~ 50

(5) 51 ~ 60 (6) 60세 이상

당신의 연령은? ( ) 세

4. 도수분포표와 막대그래프, 히스토그램

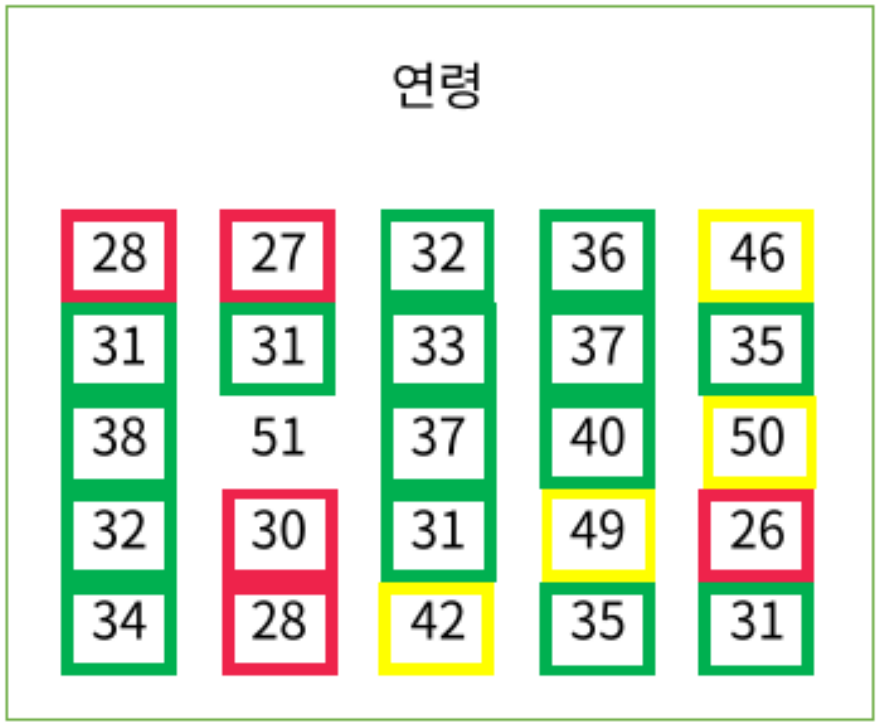
4.1 도수분포표

- 수집된 자료를 쉽게 이해할 수 있도록 일목요연하게 정리된 표로, **특정 항목 또는 특정 범위에 속하는 빈도수를 나타낸 표**

범위	빈도수

4. 도수분포표와 막대그래프, 히스토그램

4.1 도수분포표



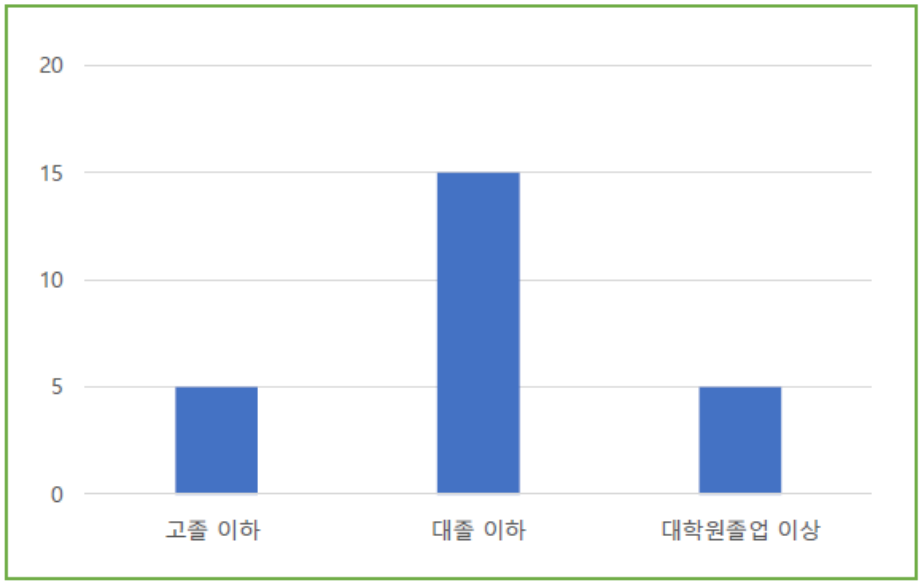
범위	빈도수
20세 초과 ~ 30세 이하	5
30세 초과 ~ 40세 이하	15
40세 초과 ~ 50세 이하	4
50세 초과 ~ 60세 이하	1
합계	25

4. 도수분포표와 막대그래프, 히스토그램

4.2 막대그래프

❖ **비연속형 변수**(명목척도 및 서열척도)에 사용되는 그래프로, **각 항목에 속하는 빈도수를 나타낸 그래프**

최종학력	빈도수
고졸이하	5
대졸이하	15
대학원졸업 이상	5
합계	25

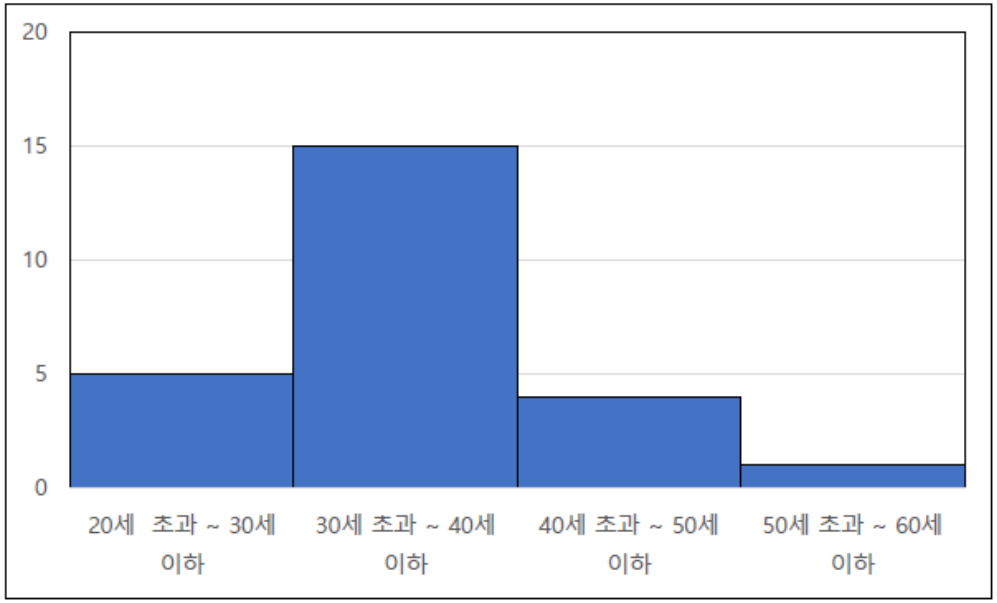


4. 도수분포표와 막대그래프, 히스토그램

4.3 히스토그램

- **연속형 변수**(등간척도 및 서열척도)에 사용되는 그래프로, **일정 범위에 속하는 빈도수를 나타낸 그래프**

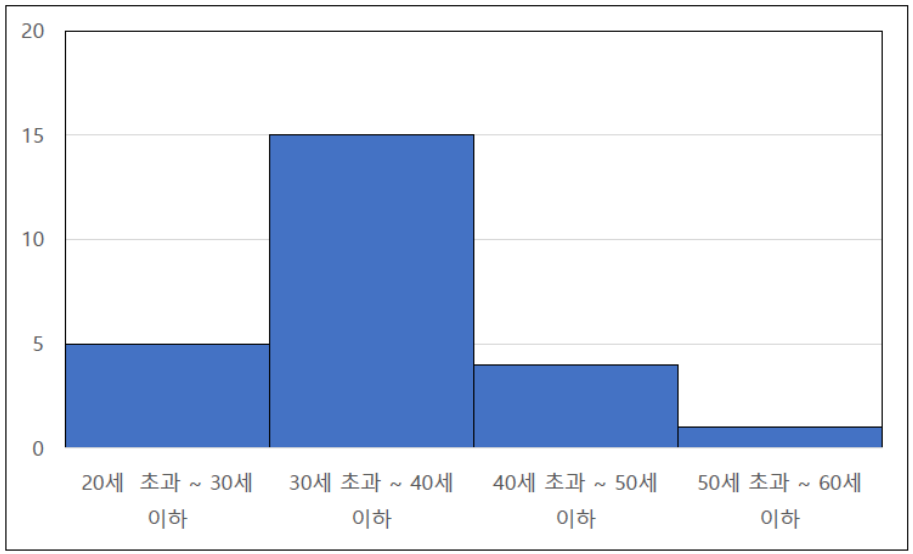
범위	빈도수
20세 초과 ~ 30세 이하	5
30세 초과 ~ 40세 이하	15
40세 초과 ~ 50세 이하	4
50세 초과 ~ 60세 이하	1
합계	25

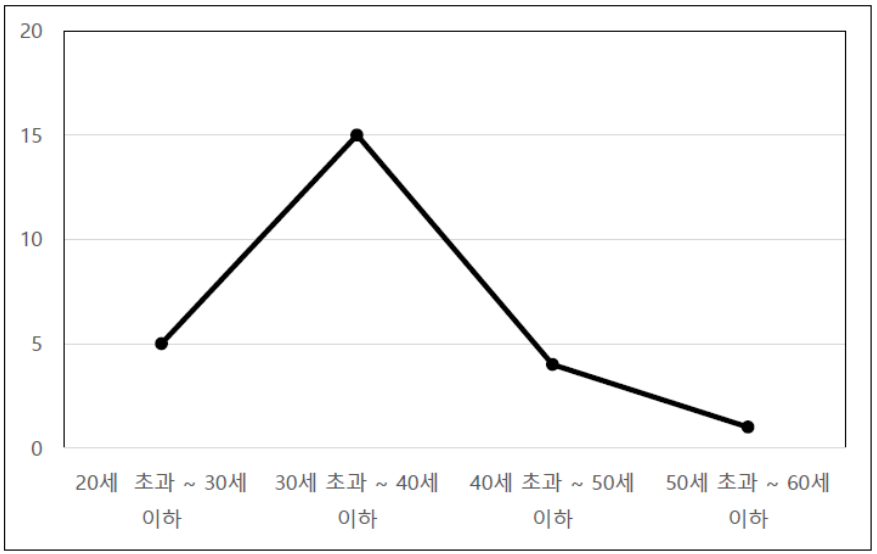


4. 도수분포표와 막대그래프, 히스토그램

4.4 선그래프

- 히스토그램의 **끝 부분을 선으로 연결한 그래프**





5. 공분산과 상관계수

5.1 공분산

- 두 변수가 함께 **각자의 평균으로부터 멀어지는 정도**
- 한 변수가 자신의 평균으로부터 멀어질 때 다른 변수가 자신의 평균으로부터 멀어지는 정도를 의미

$$\mathrm{cov}(A,B) = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{n}$$

5. 공분산과 상관계수

5.2 상관계수

- 두 변수 간의 관계로, 하나의 변수가 변화함에 따라 다른 변수가 변화하는 정도를 의미

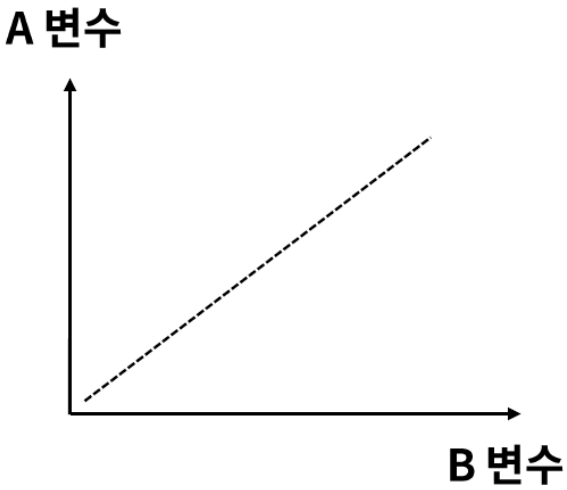
$$r_{AB} = \frac{\mathrm{cov}(A, B)}{s_a \times s_b}$$

- -1 에서 1사이의 범위를 가짐

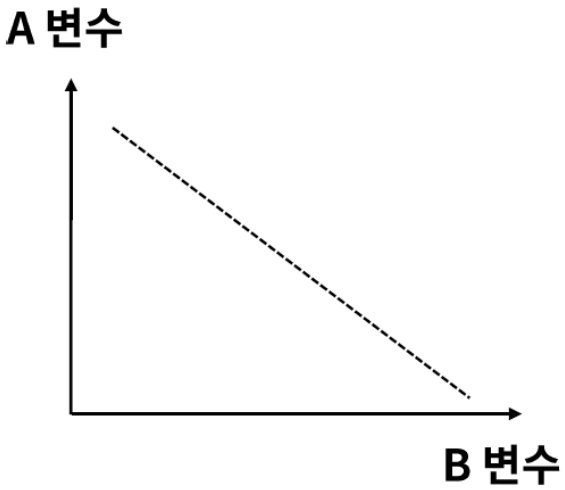
5. 공분산과 상관계수

5.2 상관계수

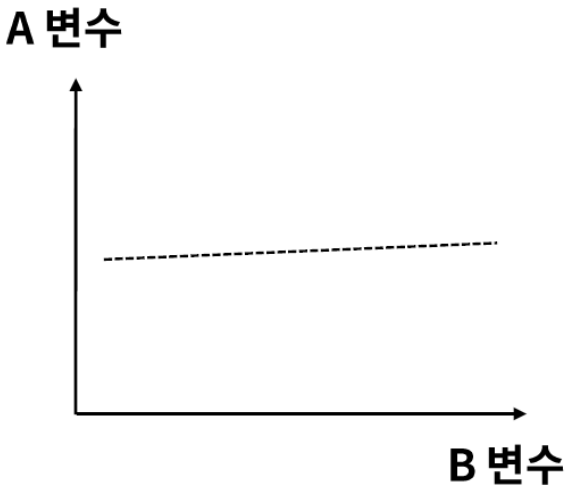
❖ 상관계수 예시



양(+)의 상관관계



음(-)의 상관관계



무의미한 상관관계

5. 공분산과 상관계수

5.2 상관계수

❖ **상관계수 예시**

변수	평균	표준편차	(1)	(2)	(3)
(1) 팀 효능감	5.14	1.00	1	-0.28***	0.37***
(2) 팀 내 정치지각	3.43	1.15	-0.28***	1	-0.07
(3) 자기효능감	5.06	0.85	0.37***	-0.07	1

❖ *** : p < 0.05, ** : p < 0.01, *** : p < 0.001 (양측 검정)**

6. 가설과 신뢰수준/유의확률

6.1 영가설과 연구가설

❖ 영가설(H_0)

- 연구가설과는 반대되는 가설이고, 실제 분석이 이루어지는 가설

❖ 연구가설(H_1)

- 분석을 통해서 알아보고자 하는 내용으로 이루어진 가설

- 통계 분석에서 영가설(H_0)이 채택 시 연구가설을 기각
- 통계분석에서 영가설(H_0)이 기각 시 연구가설을 채택

6. 가설과 신뢰수준/유의확률

6.2 예시

❖ 집단 간 차이 검증

- H_0 : A 집단의 평균과 B 집단의 평균 간에는 차이가 없다.
- H_1 : A 집단의 평균과 B 집단의 평균 간에는 차이가 있다.

❖ 영향력 검증

- H_0 : A변수가 B변수에 아무런 영향을 미치지 못할 것이다.
- H_1 : A변수가 B변수에 유의미한 영향을 미칠 것이다.

6. 가설과 신뢰수준/유의확률

6.3 유의확률

- 실제로는 영가설이 참(채택)임에도 불구하고 통계분석을 통해 영가설을 거짓(기각)으로 판단할 가능성 (p-value)

- 즉, 연구결과가 실제 현상을 반영하지 못할 가능성

- 예를 들어, 영가설(H_0)로 'A 집단의 평균과 B 집단의 평균 간에는 차이가 없다.' 라고 설정할 경우, 실제 두 집단 간에 차이 없음에도 차이가 있다고 결론내릴 가능성

6. 가설과 신뢰수준/유의확률

6.4 신뢰수준

- 실제로는 영가설이 참(채택)이고 통계분석을 통해서도 영가설을 참(채택)으로 판단할 가능성

- 즉, 실제 현상에서 발생하지 않는 연구가설을 기각할 가능성

- 예를 들어, 영가설(H_0)로 'A 집단의 평균과 B 집단의 평균 간에는 차이가 없다.' 라고 설정할 경우, 실제 두 집단 간에 차이 없으며, 두 집단의 차이가 없다고 결론 내릴 가능성

- 신뢰수준이 높아질수록 영가설(H_0)이 채택될 가능성이 높아지고, 반대로 연구가설(H_1)이 채택될 가능성이 낮아짐

- 즉, 신뢰수준이 높아질수록 연구가설이 실제 현상을 반영할 가능성이 상승

6. 가설과 신뢰수준/유의확률

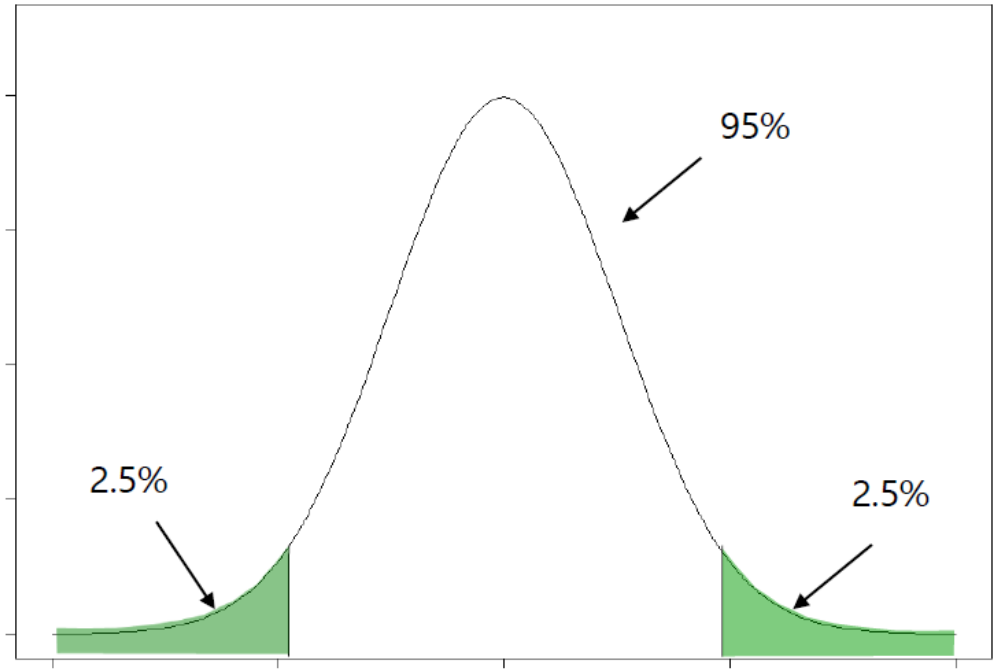
6.5 가설의 판단기준

- 95% 신뢰수준 (유의확률 0.05 미만): *
- 99% 신뢰수준 (유의확률 0.01 미만): **
- 99.9% 신뢰수준 (유의확률 0.001 미만): ***
- 90% 신뢰수준 (유의확률 0.1 미만): †

7. 양측 검증과 단측 검증

7.1 양측 검증

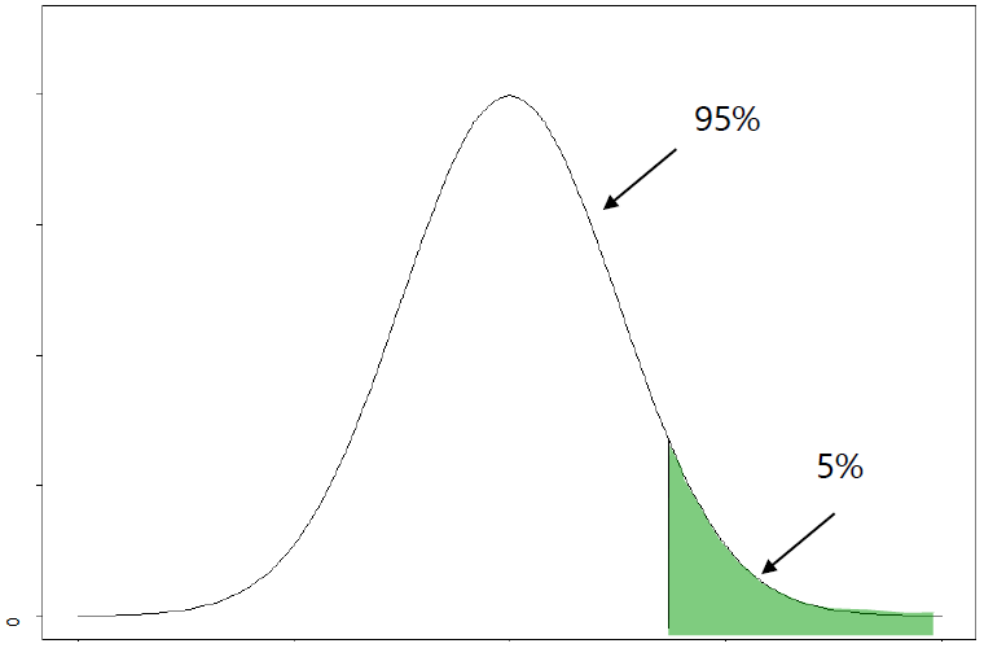
❖ **방향성을 고려하지 않은 채로 연구가설(H_1)을 설정할 때 사용하는 검증 방법**



7. 양측 검증과 단측 검증

7.2 단측 검증

❖ **방향성을 고려**하여 연구가설을 설정할 때 사용하는 검증 방법



7. 양측 검증과 단측 검증

7.3 양측검증과 단측검증의 예시

❖ 양측 검증

- A집단의 평균과 B집단의 평균 간에는 차이가 있을 것이다.
- A변수가 B변수에 미치는 영향의 크기는 '0'이 아니다.

❖ 단측 검증

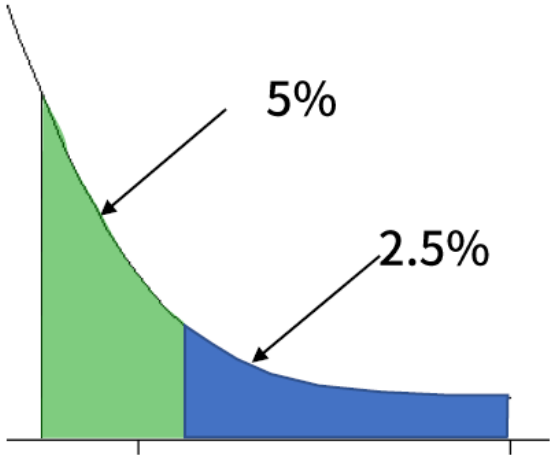
- A 집단의 평균보다 B 집단의 평균이 클 것이다. (또는 작을 것이다)
- A변수가 B변수에 미치는 영향의 크기는 '0'보다 클 것이다. (또는 작을 것이다)

7. 양측 검증과 단측 검증

7.4 연구가설(H_1)의 채택 가능성

- 양측 검증 보다는 단측 검증일 경우에 연구가설(H_1)이 채택될 가능성이 높다.

- 95% 신뢰수준의 단측 검증 = 90% 신뢰수준의 양측 검증



8. t-분석

8.1 t-분석방법

- **독립변수가 비연속형 변수**(즉, 명목척도나 서열척도)이고, **종속변수가 연속형 변수**(즉, 등간척도나 비율척도)일 때 사용하는 분석방법으로, **독립변수의 집단이 2개 이하일 때 사용하는 분석방법**

- **t-분포를 사용하여 분석**

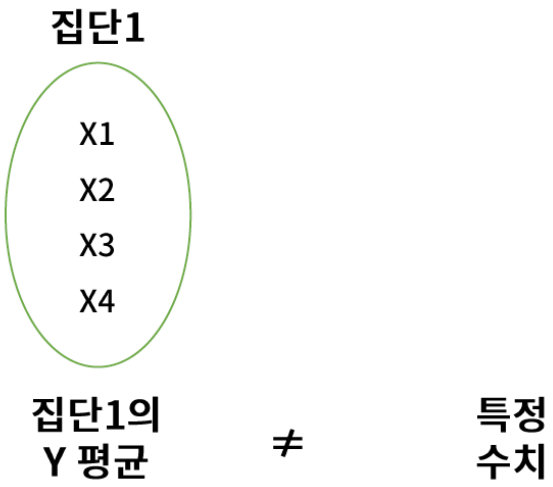
8. t-분석

8.2 t-분석의 종류(일표본)

❖ 일표본 t-분석

- 하나의 모집단에서 표본을 추출할 때 사용되는 분석으로 표본의 평균이 예측한 특정 수치와 같은 지 아니면 다른 지를 검증하는 방법

- H_0 : 국내 중학생의 평균 키는 170cm이다.
- H_1(양측 검증) : 국내 중학생의 평균 키는 170cm가 아니다.
- H_1(단측 검증) : 국내 중학생의 평균 키는 170cm보다 크다.



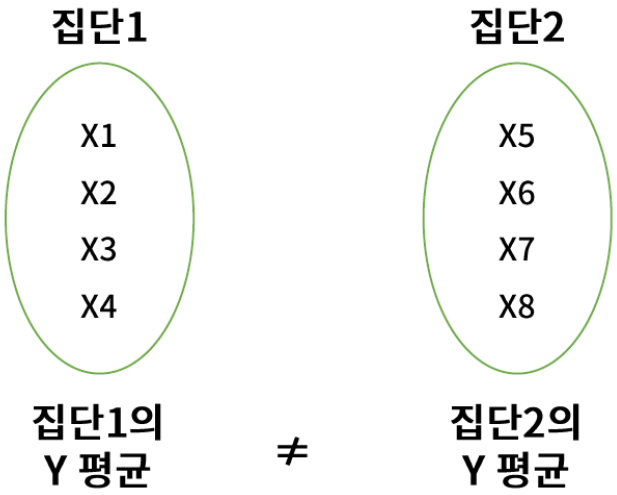
8. t-분석

8.3 t-분석의 종류(독립표본)

❖ 독립표본 t-분석

- 두개의 모집단에서 각각의 표본을 추출할 때 사용되는 분석으로 두 집단의 표본들의 평균이 서로 같은 지 다른 지를 검증하는 방법

- H_0 : A 집단의 평균과 B 집단의 평균은 같다.
- H_1(양측 검증) : A집 단의 평균과 B집 단의 평균은 다르다.
- H_1(단측 검증) : A집단의 평균은 B집단의 평균보다 크다.



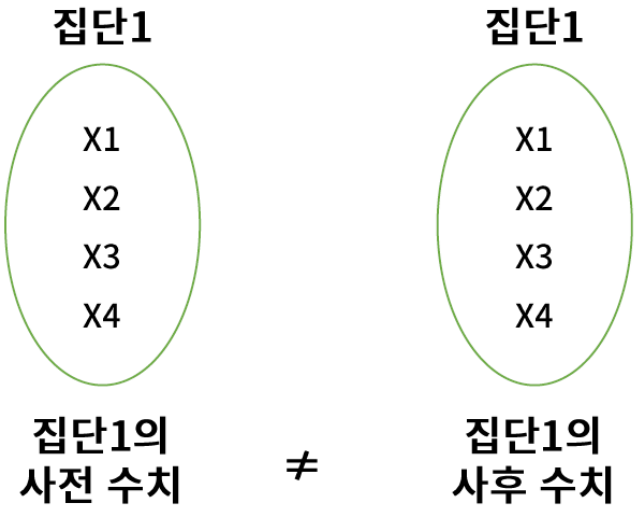
8. t-분석

8.3 t-분석의 종류(대응표본)

❖ 대응표본 t-분석

- 하나의 모집 단에서 표본을 추출하지만, 같은 표본에게 두 번의 측정이 이루어질 때 사용

- H_0 : 사전 수치와 사후 수치는 같다.
- H_1(양측 검증) : 사전 수치와 사후 수치는 다르다.
- H_1(단측 검증) : 사전 수치보다 사후 수치가 더 크다.



9. 분산분석(ANOVA)

9.1 분산분석 방법

- **독립변수가 비연속형 변수**(즉, 명목척도나 서열척도)이고, **종속변수가 연속형 변수**(즉, 등간척도나 비율척도)일 때 사용하는 분석방법으로, **독립변수의 집단이 3개 이상일 때 사용하는 분석방법**

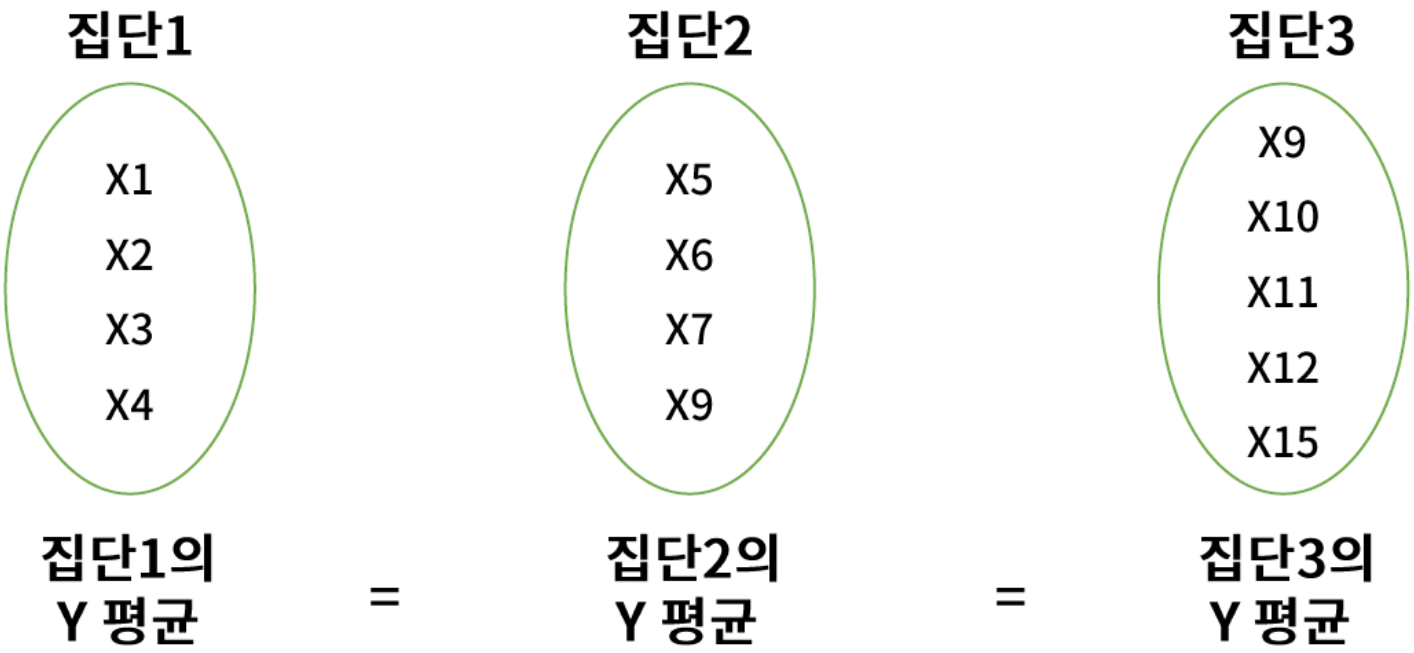
- **F-분포를 사용하여 분석**

9. 분산분석(ANOVA)

9.2 분산분석 대표 가설

- H0: 집단들의 **평균은 모두 같다**.

- H1: 집단들의 **평균은 서로 다르다**.



9. 분산분석(ANOVA)

9.3 분산분석의 원리

- **집단 간 분산과 집단 내 분산을 통해 분석**
  - 집단간 분산 > 집단내 분산: 집단간 차이가 있음
  - 집단간 분산 < 집단 내 분산: 집단간 차이가 크지 않음
  - 실제 분석은 (집단 간 분산)/(집단 내 분산)을 활용

- **사후분석: 어떠한 집단들 간에 평균 차이가 발생하는 지를 알아보기 위한 분석방법**

9. 분산분석(ANOVA)

9.4 분산분석의 종류

- 1-way ANOVA : 독립변수 1개, 종속변수1개
- 2-wayANOVA : 독립변수 2개, 종속변수1개
- 3-wayANOVA : 독립변수 3개, 종속변수1개
- ANCOVA : 독립변수 1개, 종속변수 1개, 통제변수 1개 이상
- MANOVA : 독립변수 1개, 종속변수 2개 이상
- MANCOVA : 독립변수 1개, 종속변수 2개 이상, 통제변수 1개 이상

10. 회귀분석

10.1 회귀분석 방법

- **독립변수와 종속변수가 모두 연속형 변수**(즉, 등간척도나 비율척도)일 때 사용하는 분석방법

- 추정방식은 OLS(Ordinary least square)로 이루어지는데, 이는 오차의 제곱을 최소화하는 직선이라는 의미

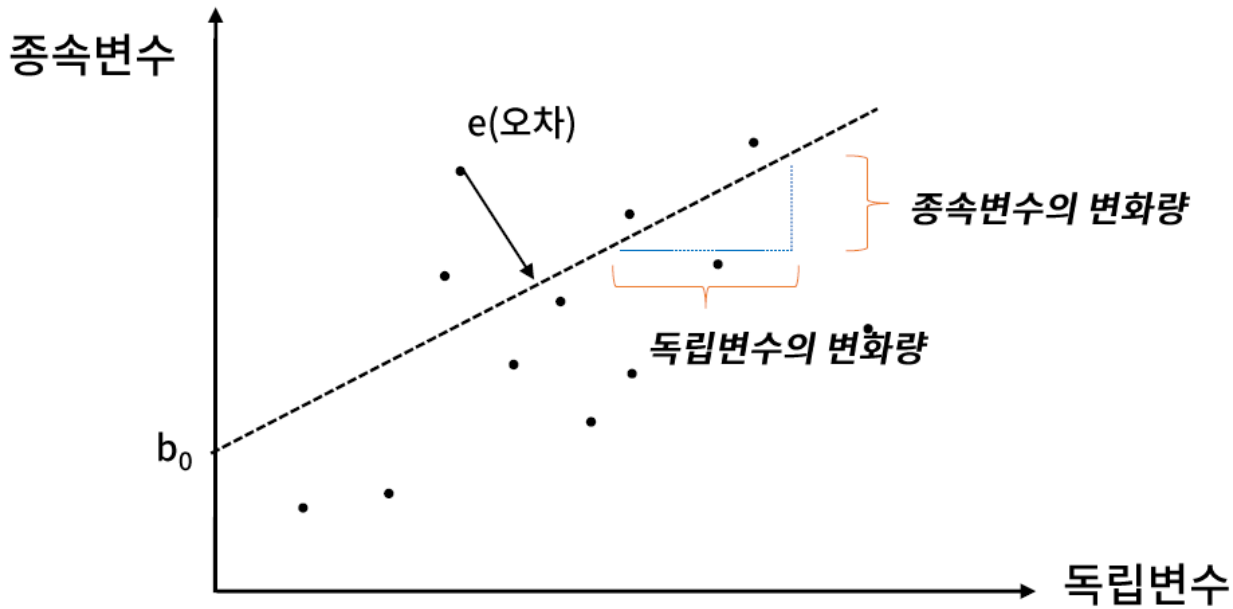
10. 회귀분석

10.2 회귀분석의 원리

$$y = b_0 + b_1 \cdot x + e$$

Y : 종속변수, X : 독립변수

b_0 : 절편, b_1: 기울기, e : 오차



10. 회귀분석

10.3 대표 가설

- H0 : 독립변수가 종속변수에 미치는 영향의 크기는 '0'이다.
- H1(양측 검증) : 독립변수가 종속변수에 미치는 영향의 크기는 '0'이 아니다.
- H1(단측 검증) : 독립변수가 종속변수에 미치는 영향의 크기는 '0' 보다 크다.

10. 회귀분석

10.4 회귀분석의 특징

- 회귀분석에서는 **여러 개의 독립변수를 포함**하는 것이 가능

- 여러 독립변수들을 포함하는 경우에는 **서로 통제되어 자신의 독자적인 영향력**으로 계산

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4$$

10. 회귀분석

10.5 설명량(R^2)

- 독립변수들에 의해서 **설명되어지는 종속변수의 분산**
- R^2가 증가할수록 회귀식에서 설명되어지지 못하는 오차는 감소
- 증가된 설명량(ΔR^2)을 이용해서 독립변수의 포함 여부를 결정

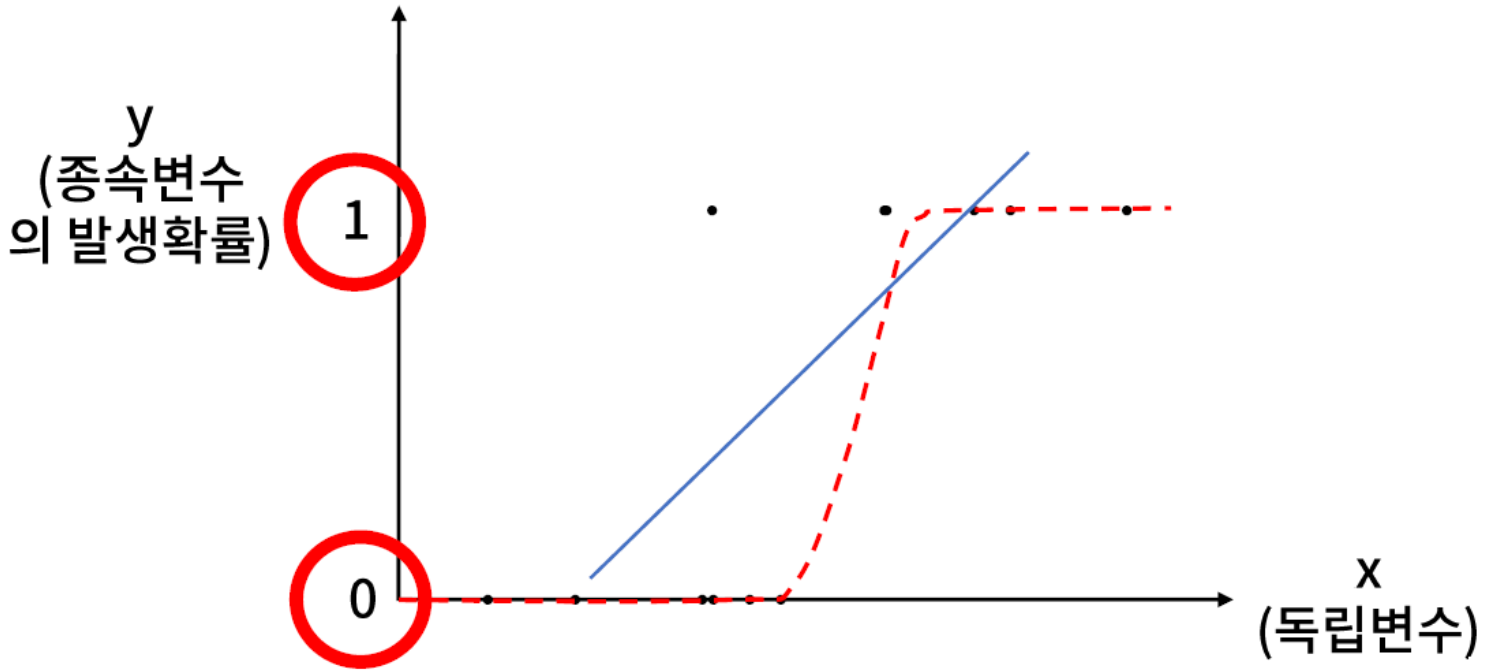
$$y_1 = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3$$
$$y_2 = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4$$

$$\left.\begin{matrix} R_1^2 \\ R_2^2 \end{matrix}\right\} \Delta R_2^2$$

11. 로지스틱 회귀분석

11.1 로지스틱 회귀분석 방법

- 독립변수가 연속형 변수이지만, 종속변수가 비연속형 변수(특히, 이분형 변수)인 경우에는 로지스틱 회귀분석을 사용하는 분석방법



11. 로지스틱 회귀분석

11.2 로지스틱 회귀분석의 원리

❖ $$Odd\ Ratio = \frac{p}{1-p}$$

➔ 특정 사건이 발생할 확률과 발생하지 않을 확률 간의 비율

❖ 로지스틱 회귀식

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \cdot x$$

➔ 회귀분석에서 종속변수(y)를 Odd 비에 자연로그를 취한 값으로 대체

❖ $b_1 > 0$: x가 증가할수록 특정 사건이 발생하지 않을 확률보다 발생할 확률이 높다는 의미

❖ $b_2 < 0$: x가 증가할수록 특정 사건이 발생할 확률보다 발생하지 않을 확률이 높다는 의미

11. 로지스틱 회귀분석

11.3 로지스틱 회귀분석 대표 가설

- H0 : 독립변수가 종속변수에 미치는 영향의 크기는 '0'이다.
- H1(양측 검증) : 독립변수가 종속변수에 미치는 영향의 크기는 '0'이 아니다.
- H1(단측 검증) : 독립변수가 종속변수에 미치는 영향의 크기는 '0' 보다 크다.

11. 로지스틱 회귀분석

11.4 모형적합도

- 모형이 적절하게 만들어졌는 지를 보여주는 지표

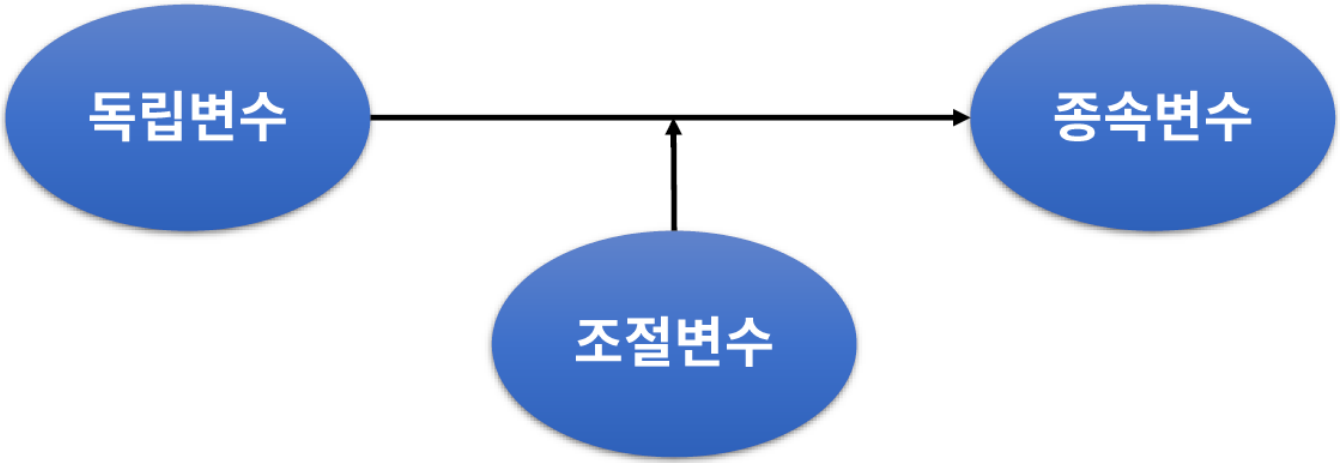
- 로지스틱 회귀분석에서는 모형에 포함된 독립변수들에 의해서 종속변수가 설명되어지는 부분

- 대표적으로 X^2-수치, -2log 우도 등이 사용

12. 조절효과와 매개효과

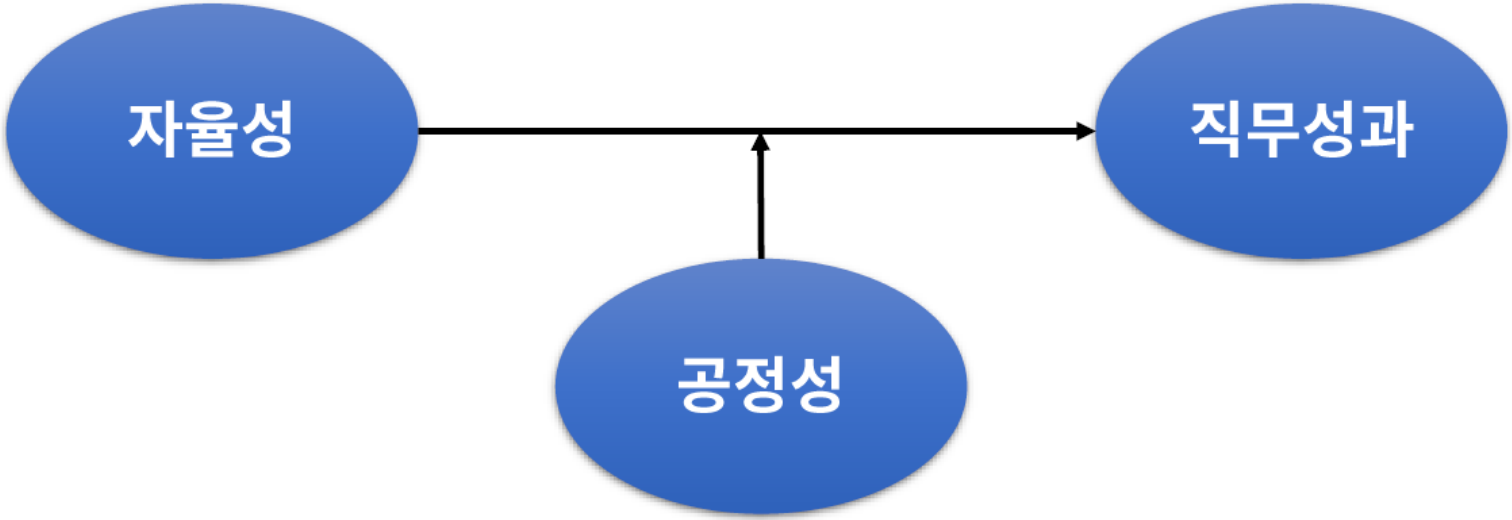
12.1 조절효과

- 독립변수가 종속변수에 미치는 영향이 조절변수에 의해서 달라지는 지를 알아보는 분석 방법



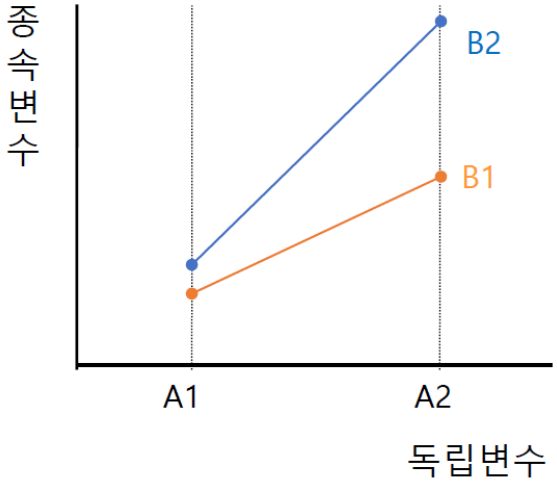
12. 조절효과와 매개효과

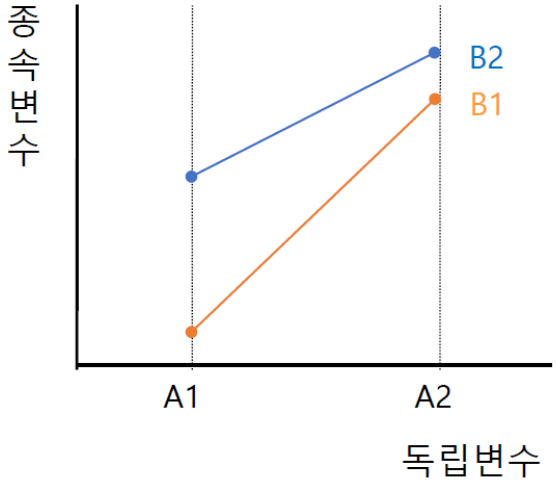
12.2 조절효과 예시

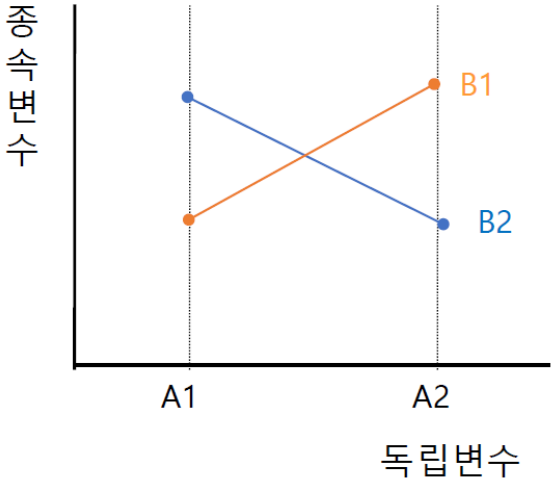


12. 조절효과와 매개효과

12.2 조절효과 유형





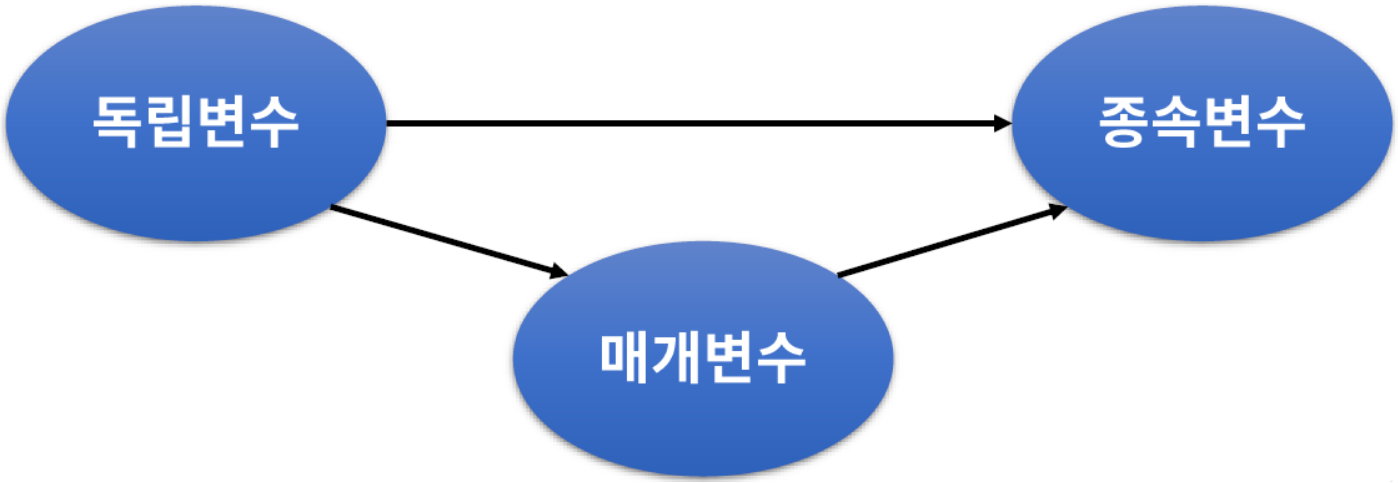


12. 조절효과와 매개효과

12.3 매개효과

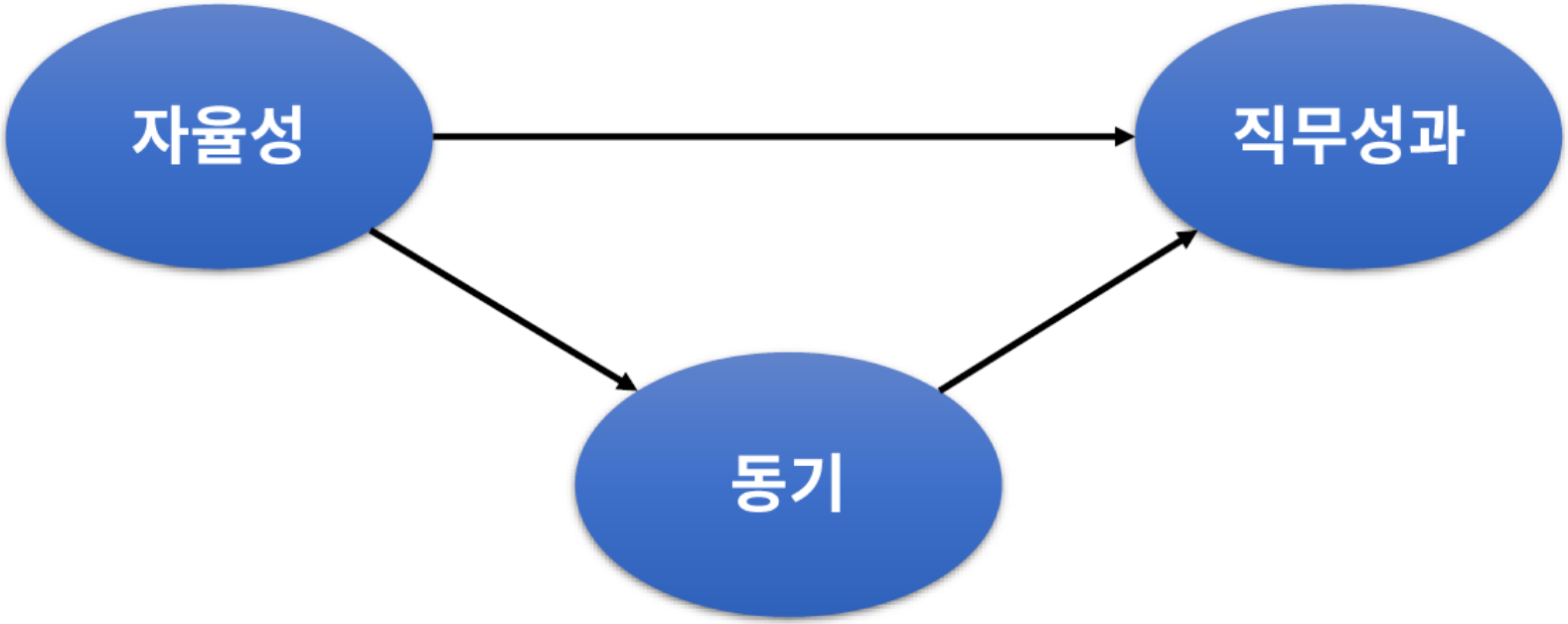
- 독립변수와 종속변수 간의 직접적인 인과관계 이외에도 매개변수를 통한 간접적인 인과관계가 존재하는 지를 알아보는 분석방법

- 총 효과 = 직접효과 + 간접효과 (또는 매개효과)



12. 조절효과와 매개효과

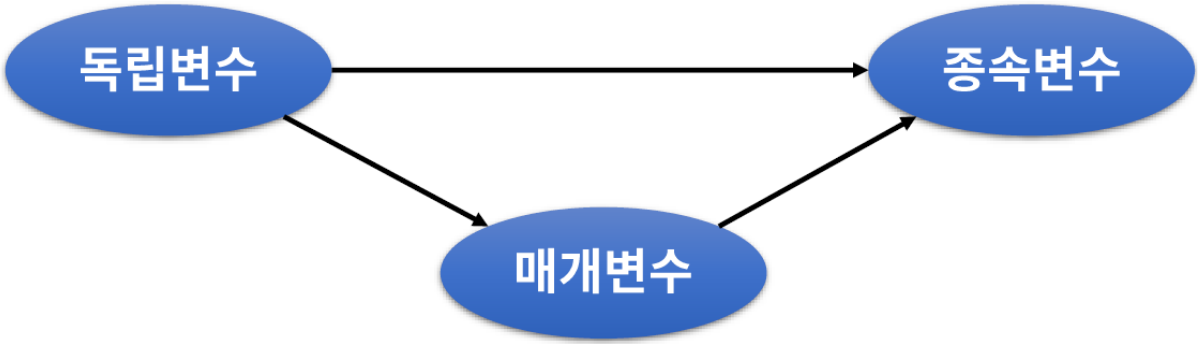
12.3 매개효과 예시



12. 조절효과와 매개효과

12.4 매개효과 유형

- 부분매개모형



- 완전매개모형

13. 구조방정식 모형

13.1 구조방정식 모형

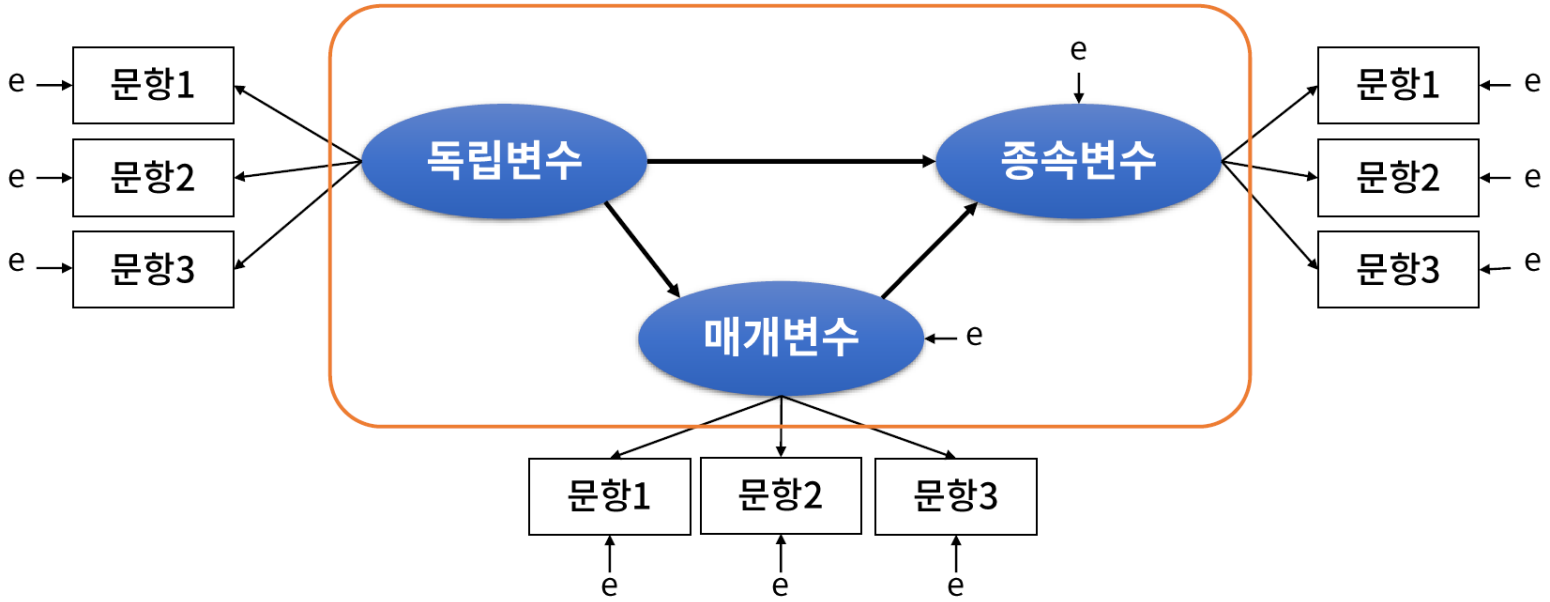
- 변수들 간의 관계를 밝히는 구조모형과 각 변수와 이를 측정하는 문항들 간의 관계를 밝히는 측정모형을 함께 고려하는 분석방법

- 확인적 요인분석과 매개효과 분석에 주로 활용

13. 구조방정식 모형

13.2 구조모형

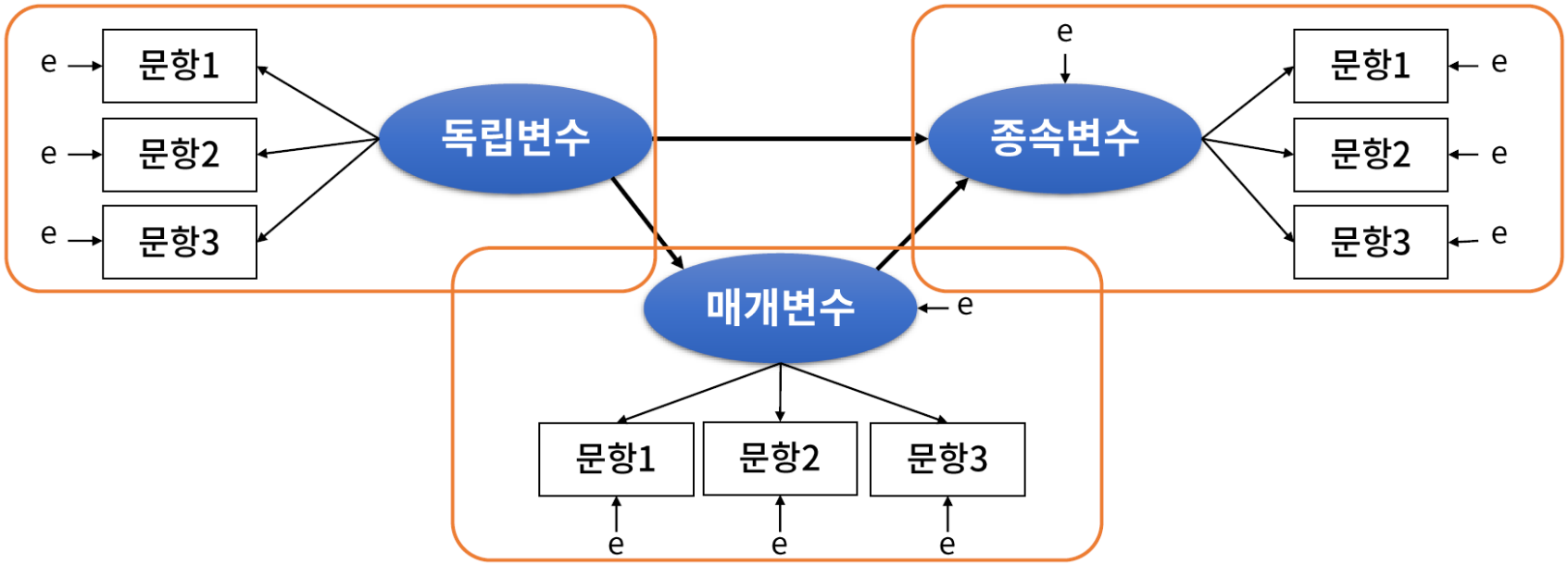
❖ 변수들 간의 관계



13. 구조방정식 모형

13.3 측정모형

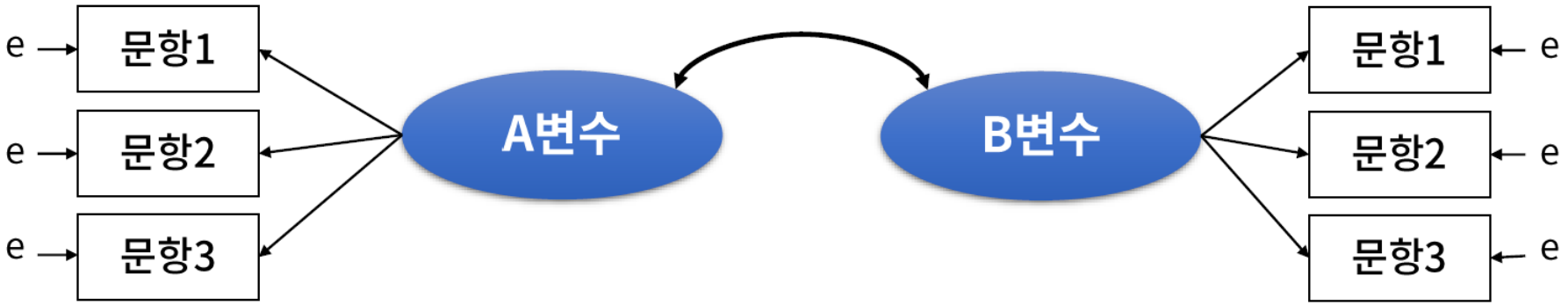
❖ 변수와 측정문항들 간의 관계



13. 구조방정식 모형

13.4 확인적 요인분석

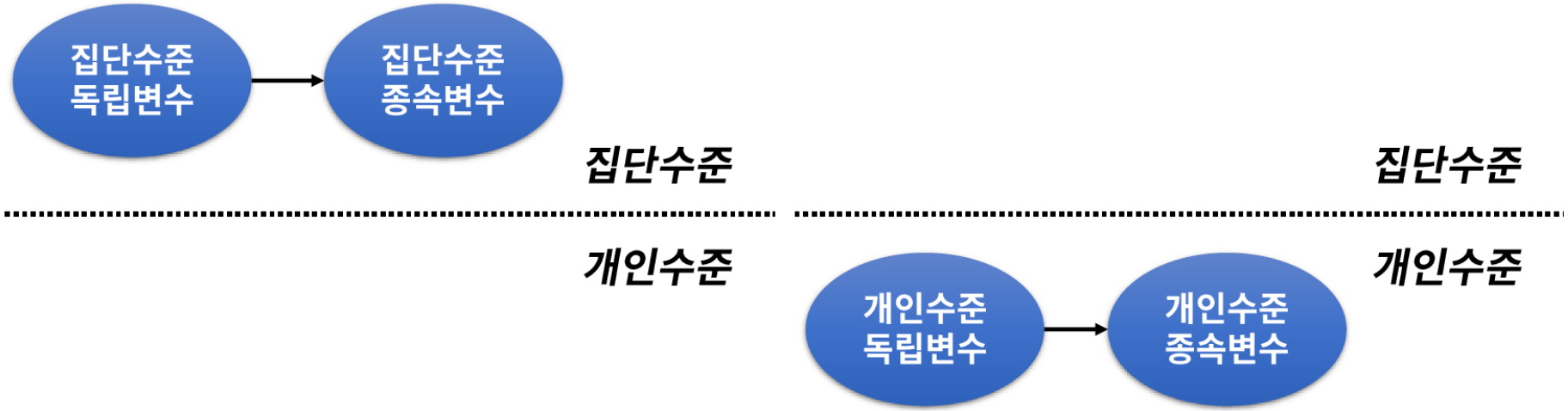
- 확인적 요인분석은 측정문항들의 타당도를 알아보기 위한 분석으로 구조방정식모형 중 측정모형만을 분석하는 방법



14. 다수준분석

14.1 단일수준분석

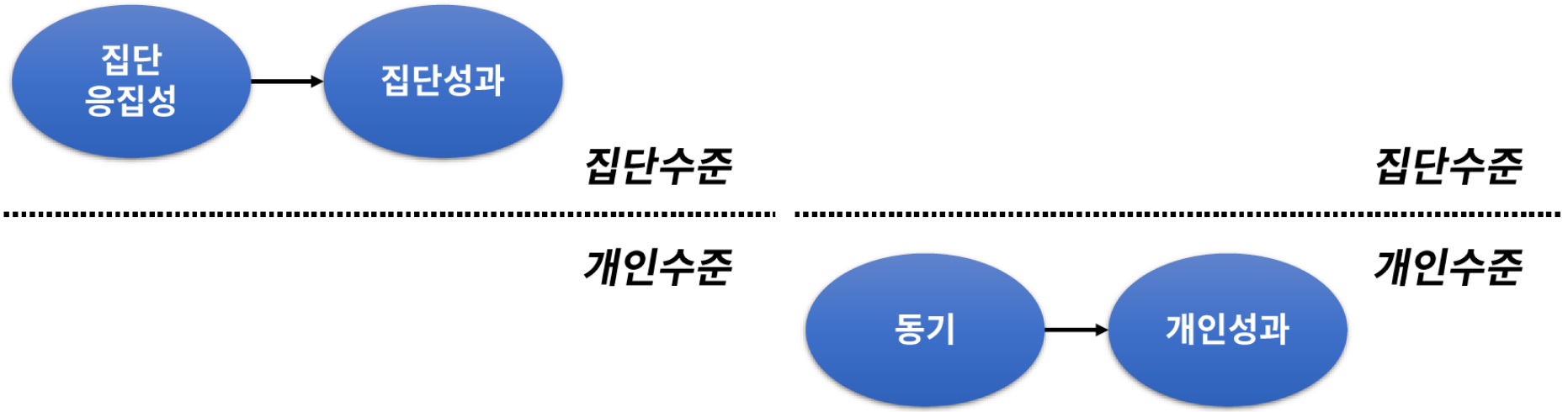
❖ 모든 변수가 하나의 수준으로 이루어진 경우



14. 다수준분석

14.2 단일수준분석 예시

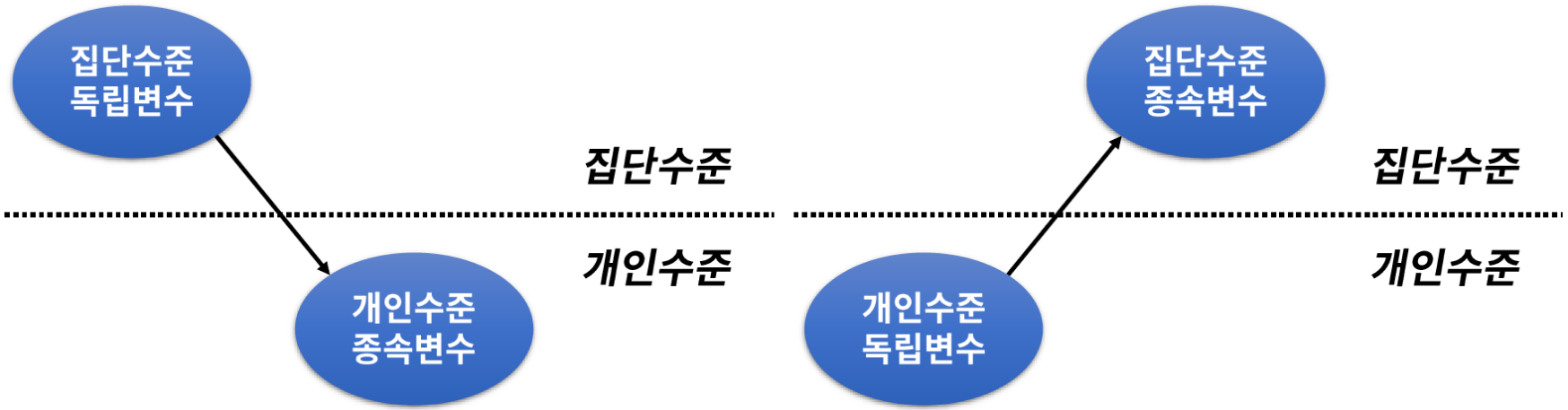
- 모든 변수가 하나의 수준으로 이루어진 경우



14. 다수준분석

14.3 다수준분석

- 독립변수와 종속변수의 수준이 다른 경우



14. 다수준분석

14.4 다수준분석 예시