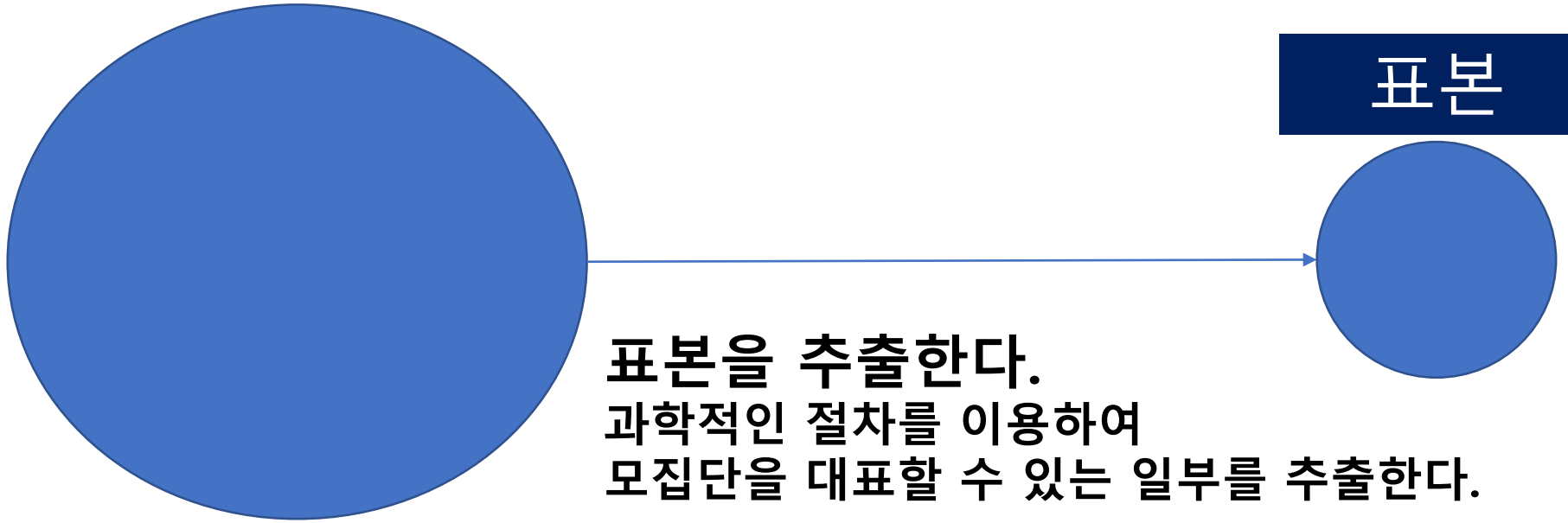
통계 기반 데이터 분석 용어 이해하기

목 차

1-1 모집단과 표본

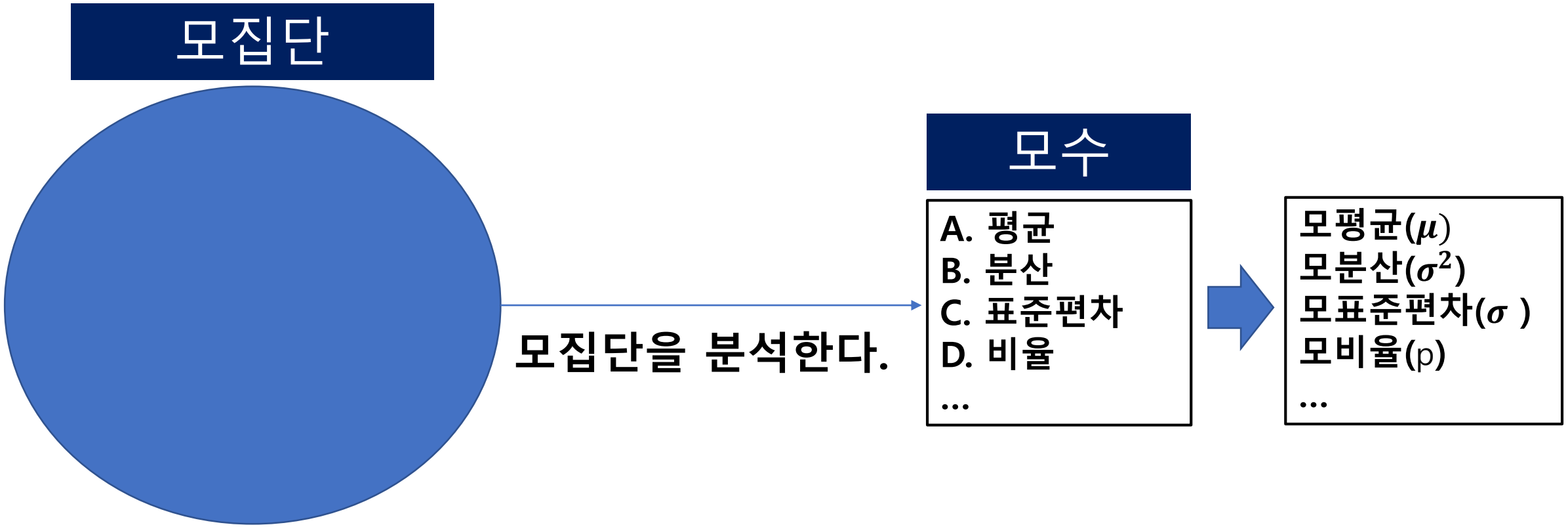
모집단 **(population)은 통계 분석 방법을 적용할 관심 대상의 전체 집합을 말한다.**



표본 **표본(sample)**
직접적인 조사 대상의 모집단의 일부

표본을 추출한다.
과학적인 절차를 이용하여 모집단을 대표할 수 있는 일부를 추출한다.

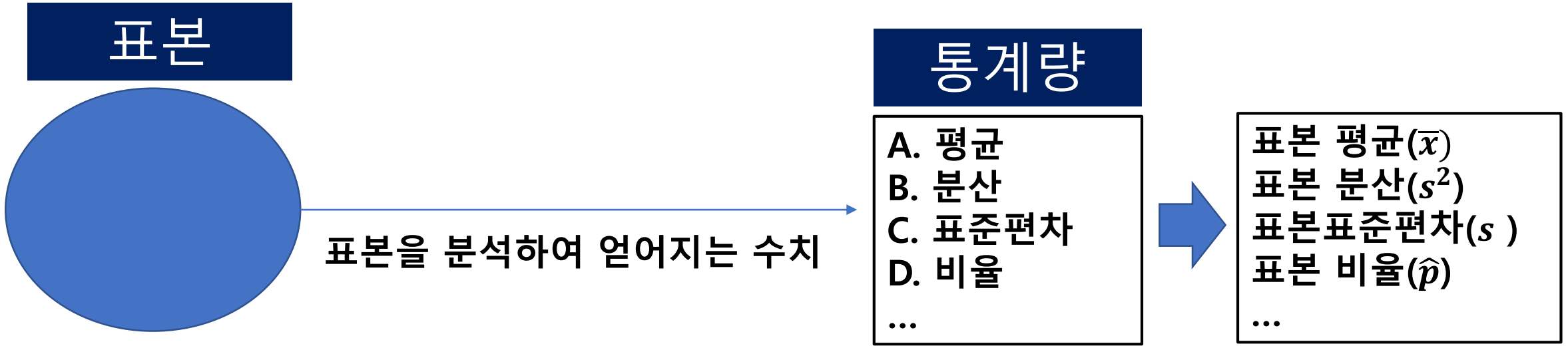
1-2 모수와 통계량



모집단을 분석한 후, 얻어지는 결과 수치를 모수(parameter)라고 한다.

1-3 모수와 통계량

표본을 분석하여 얻어지는 수치를 우리는 통계량(statistic)라 한다.

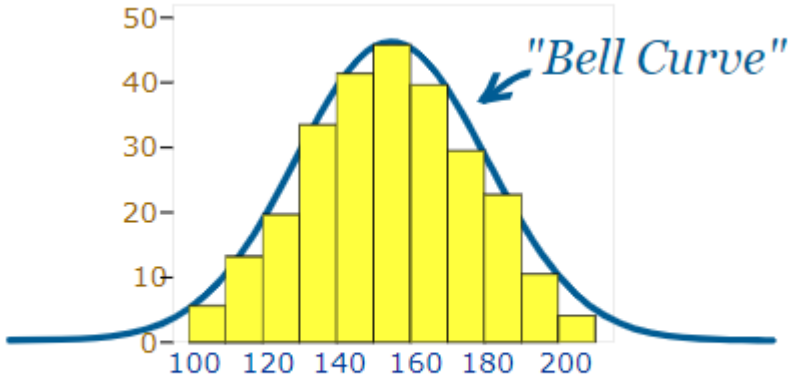


1-4 표본의 분포-정규분포

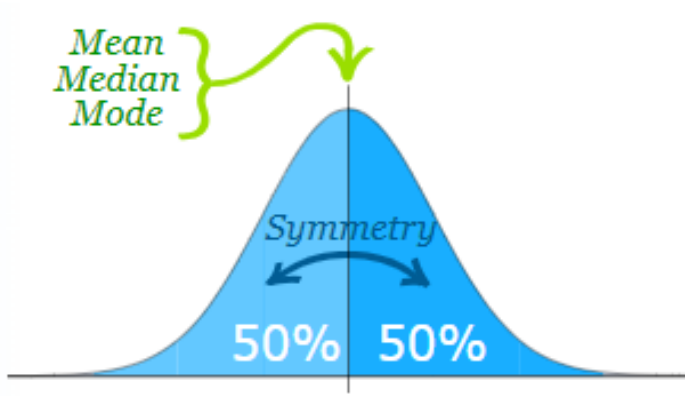
A. 정규분포는 통계학에서 가장 많이 사용되는 분포이다.

B. 평균과 분산만으로 그 특성을 모두 설명할 수 있다.

C. 정규 분포는 평균을 중심으로 좌우대칭인 종의 모양을 하고 있다.



A Normal Distribution



https://www.mathsisfun.com/data/standard-normal-distribution.html

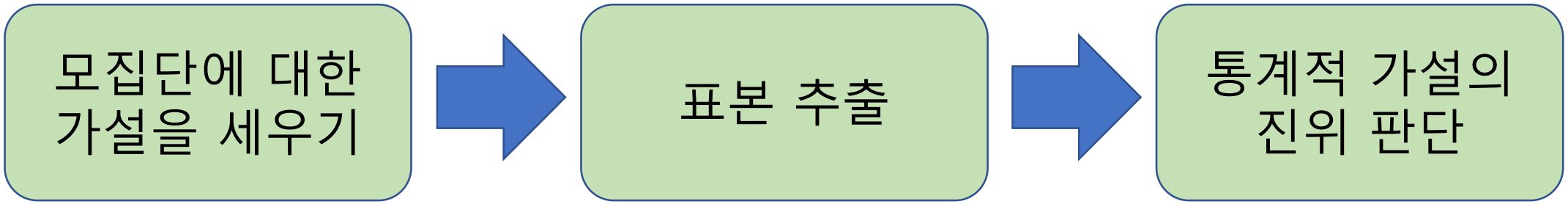
가설 검정이란 무엇일까?

1-5 가설 검정

▶ 가설 검정?

가. 모수가 어떠할 것인가에 대해 '맞다 ' 혹은 '아니다'를 판단하는 방법이다.

나. 주어진 유의수준(α) 하에서 주장이나 추측이 일정 신뢰 구간에 포함될 지의 여부를 판단하는 것.



신뢰구간? 유의수준?

1-5 가설 검정- 신뢰구간

▶ 신뢰구간(confidence interval)

가. 신뢰구간은 진정한 가치가 있음을 확신하는 범위의 값.

나. 알 수 없는 모집단의 값을 관측된 데이터의 통계량으로부터 계산된 간격 추정 유형.
(wiki 참조)

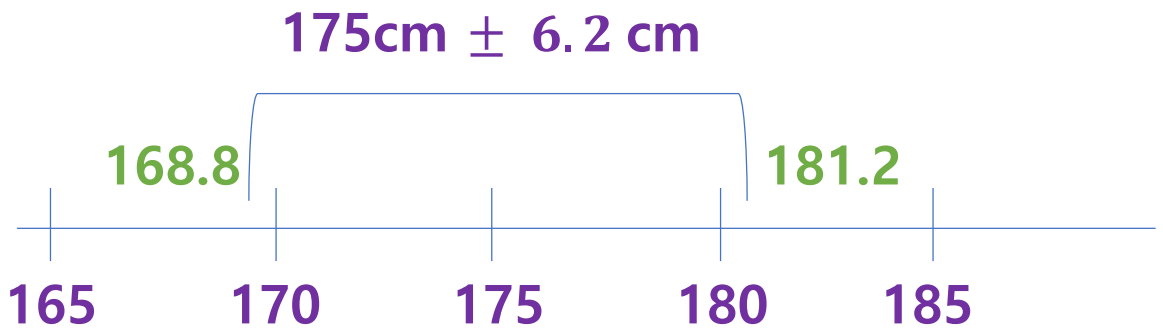
1-5 가설 검정- 신뢰구간

▶ 신뢰구간(confidence interval)

(예) 키의 평균

가. 40명을 임의로 선택 후, 평균 175cm 확인함.

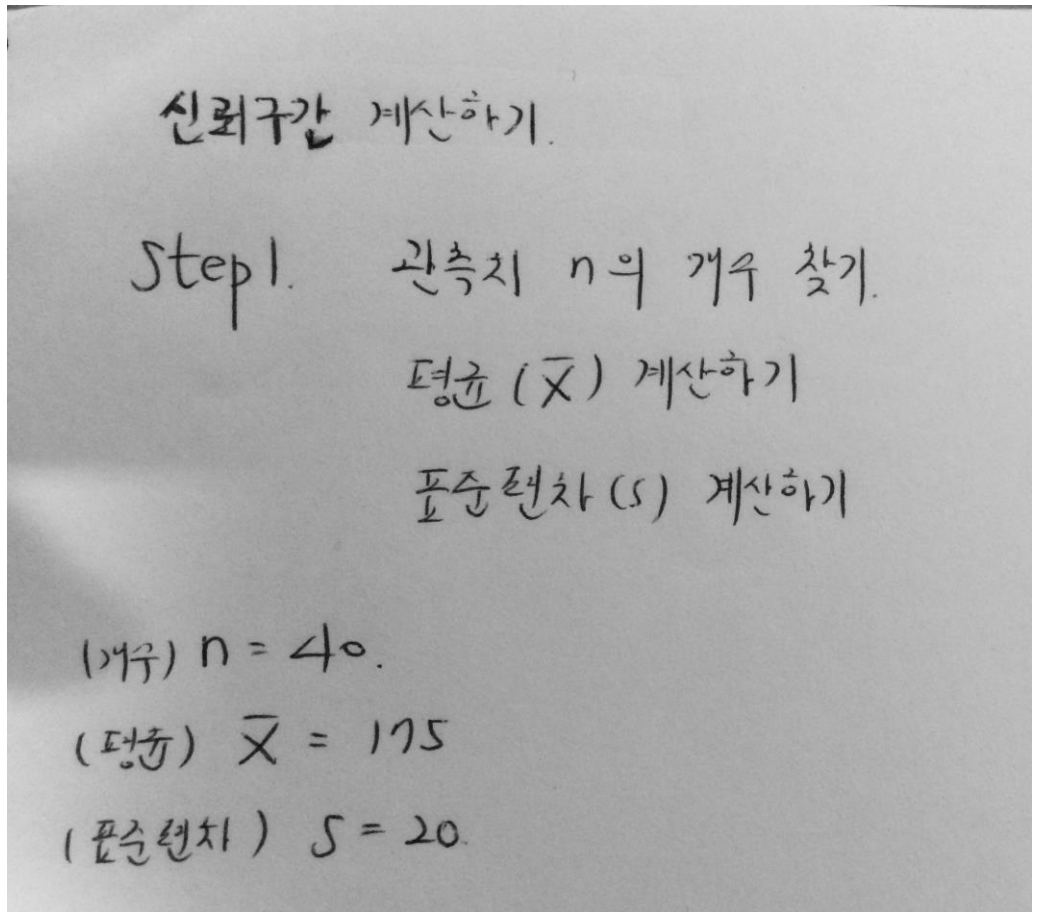
나. 또한 표준편차가 20cm임을 알았음.

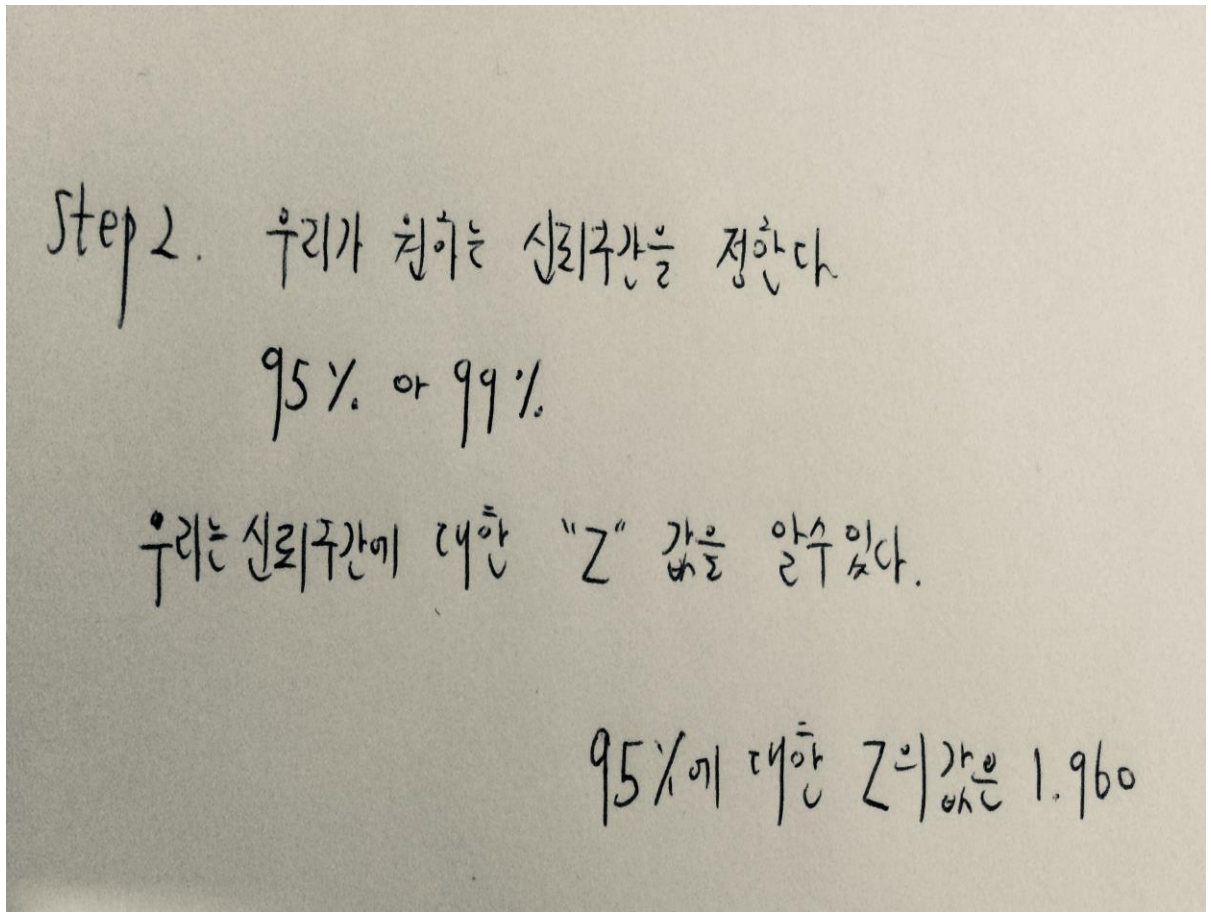


95%의 신뢰도로 신뢰구간은 168.8~181.2이다.

1-5 가설 검정- 신뢰구간

▶ 신뢰구간(confidence interval) 계산하기

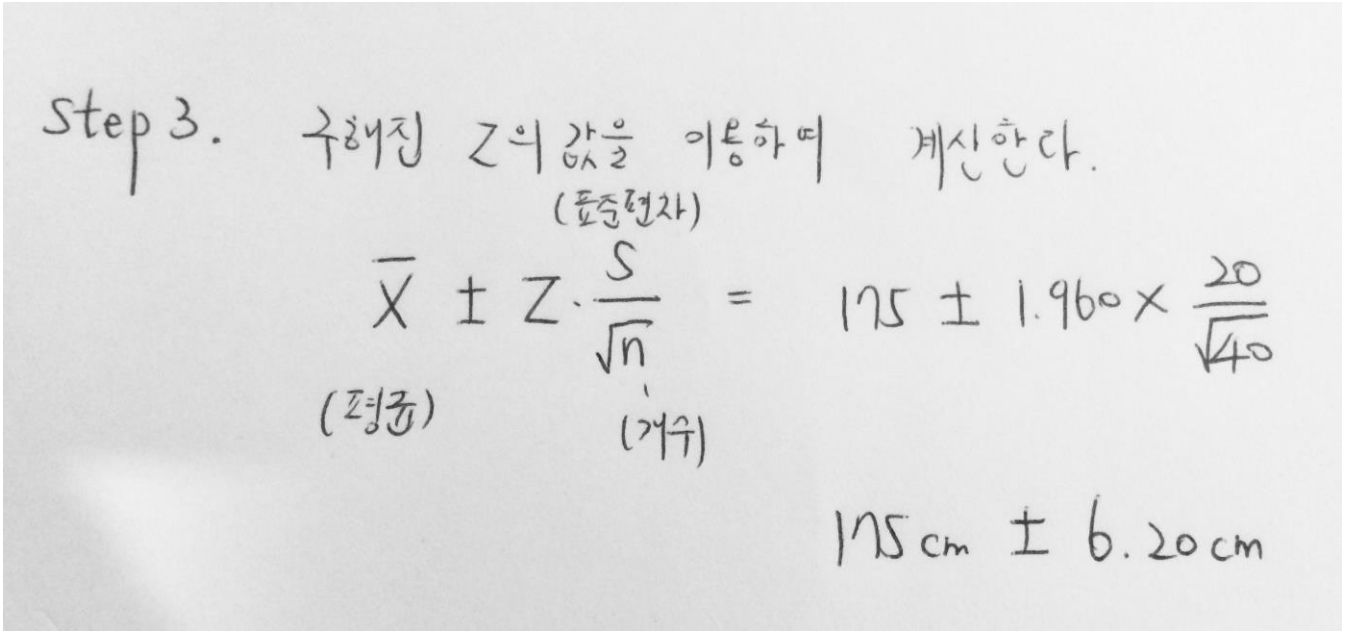




1-5 가설 검정- 신뢰구간

▶ 신뢰구간(confidence interval) 계산하기

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291



신뢰구간 계산기

https://www.mathsisfun.com/data/confidence-interval-calculator.html

1-5 가설 검정 - 유의수준

▶ 유의수준(significance level)

통계적인 가설검정에서 사용되는 기준값이다.

(가) 표기

일반적으로 유의 수준은 α로 표시한다.

(나) 신뢰수준(level of confidence)

유의 수준이 0.05 라면 신뢰수준은 $(1-\alpha)$이다. 0.95가 신뢰수준 값이다.

신뢰수준의 값이 0.95라면 귀무가설이 참일 때, 참이라고 판단하는 확률이 95%이다.

(다) 통계적으로 유의하다

가설 검정의 절차에서 유의수준 값과 유의확률 값을 비교하여 통계적

유의성을 검정한다.

유의수준은 얼마나 기준일까?

1-5 가설 검정 - 유의수준

▶ 5% 유의수준(significance level)

유의수준 5%란 통계 분석에서 제 1종 오류를 범할 확률을 5% 미만으로 제한하겠다.

▶ 유의 수준의 결정

통계학적으로 유의수준(제 1종 오류를 범할 최대허용 확률)을 어떻게 결정할지에 대한 명확한 이론은 없다.(NCS 모듈 교재)

따라서 통상적인 허용되는 수준으로 결정한다.(0.1, 0.05, 0.01 중 하나)

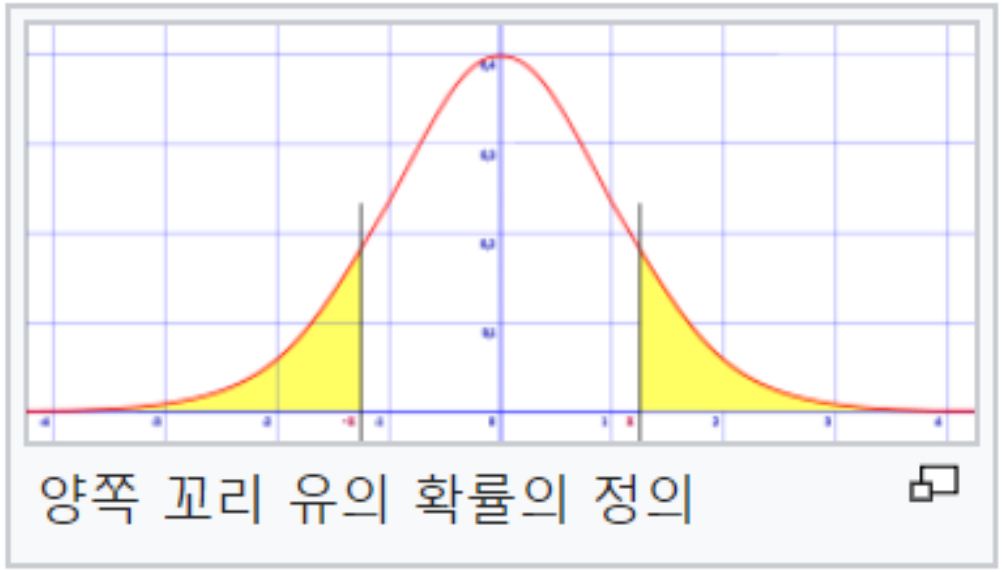
1-5 가설 검정 – 유의확률(significance probability)

▶ 유의 확률(significance probability) , p값

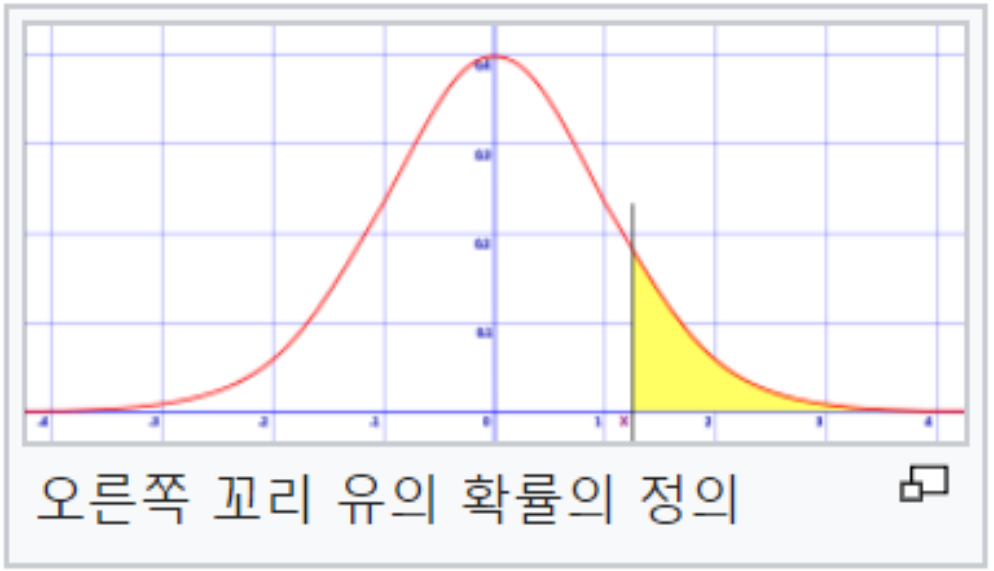
가. 귀무 가설이 맞다고 가정,

얻은 결과보다 극단적인 결과가 실제로 관측될 확률

나. 유의 확률은 실험의 표본 공간에서 정의되는 확률 변수로 0~1사이의 값을 가진다.



양쪽 꼬리 유의 확률의 정의



오른쪽 꼬리 유의 확률의 정의

1-5 가설 검정

▶ 왜 가설 검정이 필요한가?

빅데이터 수집

전체 데이터의 일부 표본

비지니스적으로 유의미한 지식 및 정보 추출

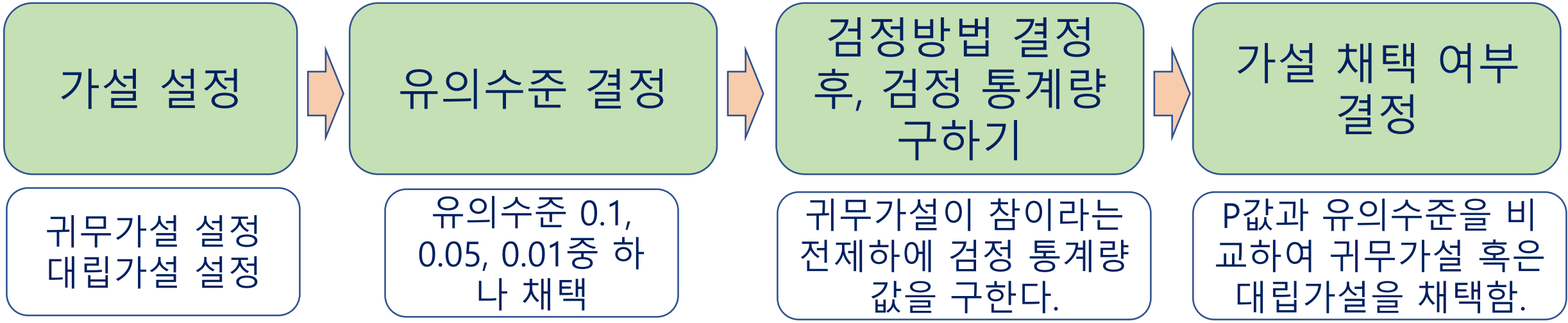
가. 다양한 통계적 방법 이용

나. 특정 변수의 미래 값을 예측, 평가하기 위한 예측 모델을 도출

다. 모델의 유의성 검증을 위해 가설 설정 및 가설검정을 시행

1-5 가설 검정 – 가설 검정 절차

▶ 가설 검정 절차



검정 통계량은 통계적 **가설 검정에 사용되는 통계량**을 말한다. 확률 표본의 함수로 표현됨.

확률 분포에 따라 검정 통계량으로 Z통계량(정규분포), t통계량(t분포), χ^2통계량(χ^2), F통계량(F분포)을 사용.

가설 검정 – 통계적 유의성 검증

유의성 검증?
통계적으로 유의하다?

* 가설 검정의 절차에서 유의수준 값과 유의확률 값을 비교하여 통계적 유의성을 검정한다.

1-5 가설 검정- 통계적 유의성

▶ 통계적 유의성

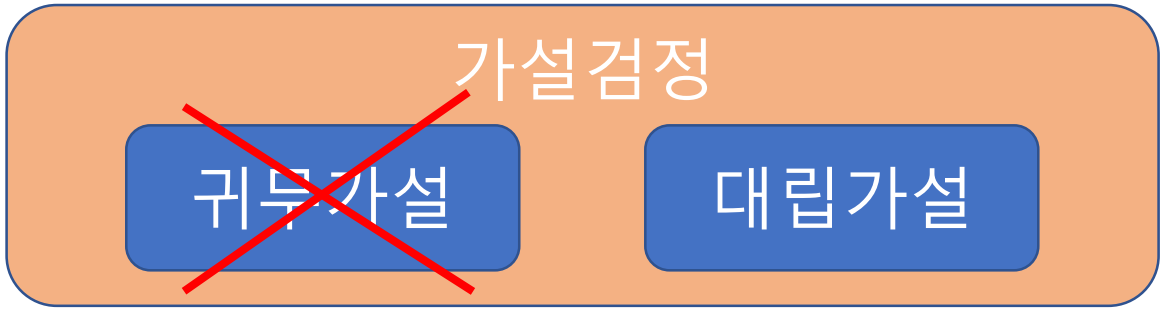
모집단에 대한 가설이 가지는 **통계적 의미**를 말한다.

▶ '통계적으로 유의하다'의 의미

확률적으로 봐서 단순한 우연이라고 생각하지 않을 정도로 **의미가 있다.**

▶ '통계적으로 유의하지 않다.'의 의미

실험 결과가 단순히 우연일 수 있다.



가설에 대해 알아보자.

1-5 가설 검정

▶ 가설(Hypothesis)이란?

가. 통계학적으로 모수는 어떠하다는 조사자의 주장이나 추측을 말한다.

나. 모집단의 특성, 특히 모수에 대한 가정 혹은 잠정적인 결론을 말한다.

▶ 귀무가설, 대립가설

기존의 주장(귀무가설)
(null hypothesis)

우리가 믿어왔으니 그대로 맞을거야?
H_0

새로운 주장(대립가설)
(alternative hypothesis)

공공연한 사실에 대립되는 가설.
영에 반대가 된다.
H_1

1-5 가설 검정

▶ 귀무 가설과 대립가설

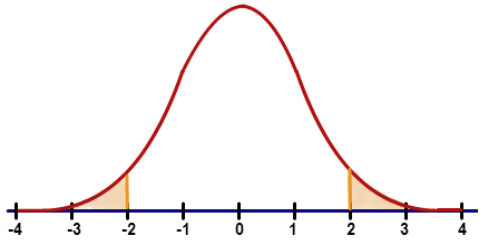
기존의 주장(귀무가설) (null hypothesis)	새로운 주장(대립가설) (alternative hypothesis)
H_0	H_1
신약 개발 제약 회사	신약 개발 제약 회사
(가) 기존 약 A와 B는 효과 차이가 없다.	(가) 기존 약 A와 B는 효과 차이가 있다.

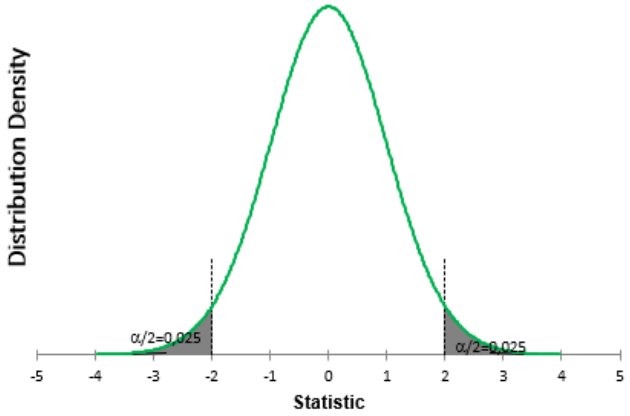
가설 검정 – 통계적 유의성 검증

1-6 양측 검정과 단측검정

▶ 양측검정(two-sided test)

(가설) 이온 음료의 용량이 300ml가 아니다.

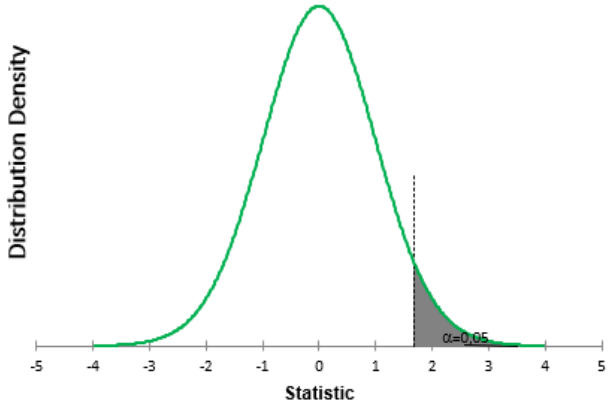




▶ 단측검정(one-sided test)

(가설) 이온 음료의 용량이 300ml보다 적다

(가설) 이온 음료의 용량이 300ml보다 많다.



https://math.tutorvista.com/statistics/two-sided-tests.html

1-7 1종 오류 vs 2종 오류

▶ 1종 오류(type I error)

* 귀무가설이 참인데 잘못하여 이를 기각하는 오류

(예) 신약이 실제로는 효과가 없는데

'실험을 했더니 효과가 있다'라는 오류

▶ 2종 오류(type II error)

* 귀무가설이 거짓인데 잘못하여 이를 채택하는 오류

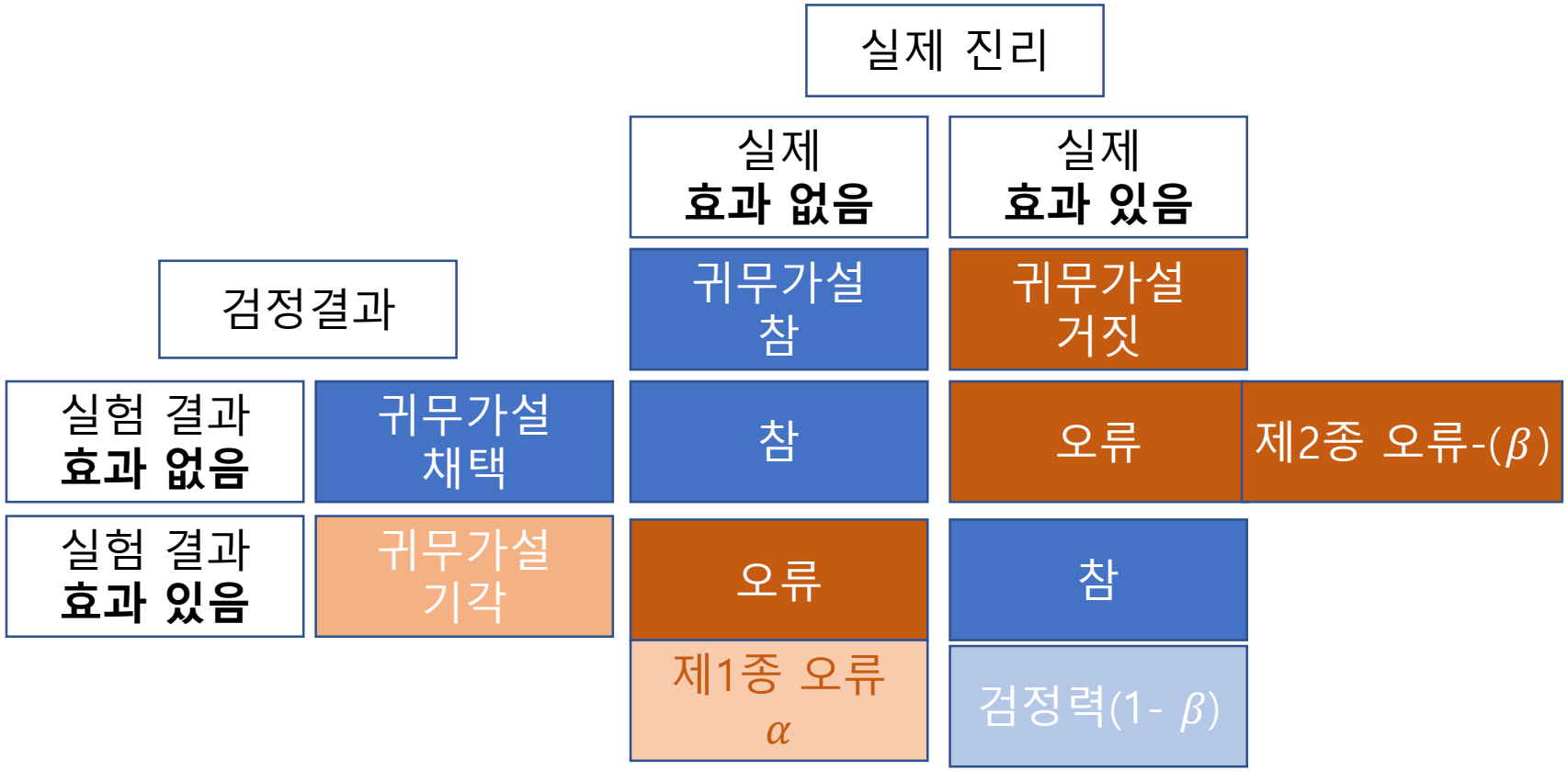
(예) 신약이 실제로는 효과가 있는데

'실험결과 효과가 없다'고 나오는 오류

일반적으로 제 2종 오류의 결과는 실험결과를 입증못했으니 발표가 안된다. 하지만 1종 오류의 경우는 실험결과 인정받고 이를 사용하게 된다면 사회적 문제가 된다.
따라서 학계에서 이 중요한 오류를 제 1종 오류라 정의하고, 1종 오류의 발생확률을 5% 미만으로 지킬 것을 권고하고 있음.(의학통계 참조)

1-7 1종 오류 vs 2종 오류

▶ 1종 오류 vs 2종 오류



1-8 검정 통계량

▶ 검정 통계량

검정 통계량은 가설 검정의 대상이 되는 모수를 추론하기 위해 사용되는 표본 통계량.

▶ 검정력(power of test)

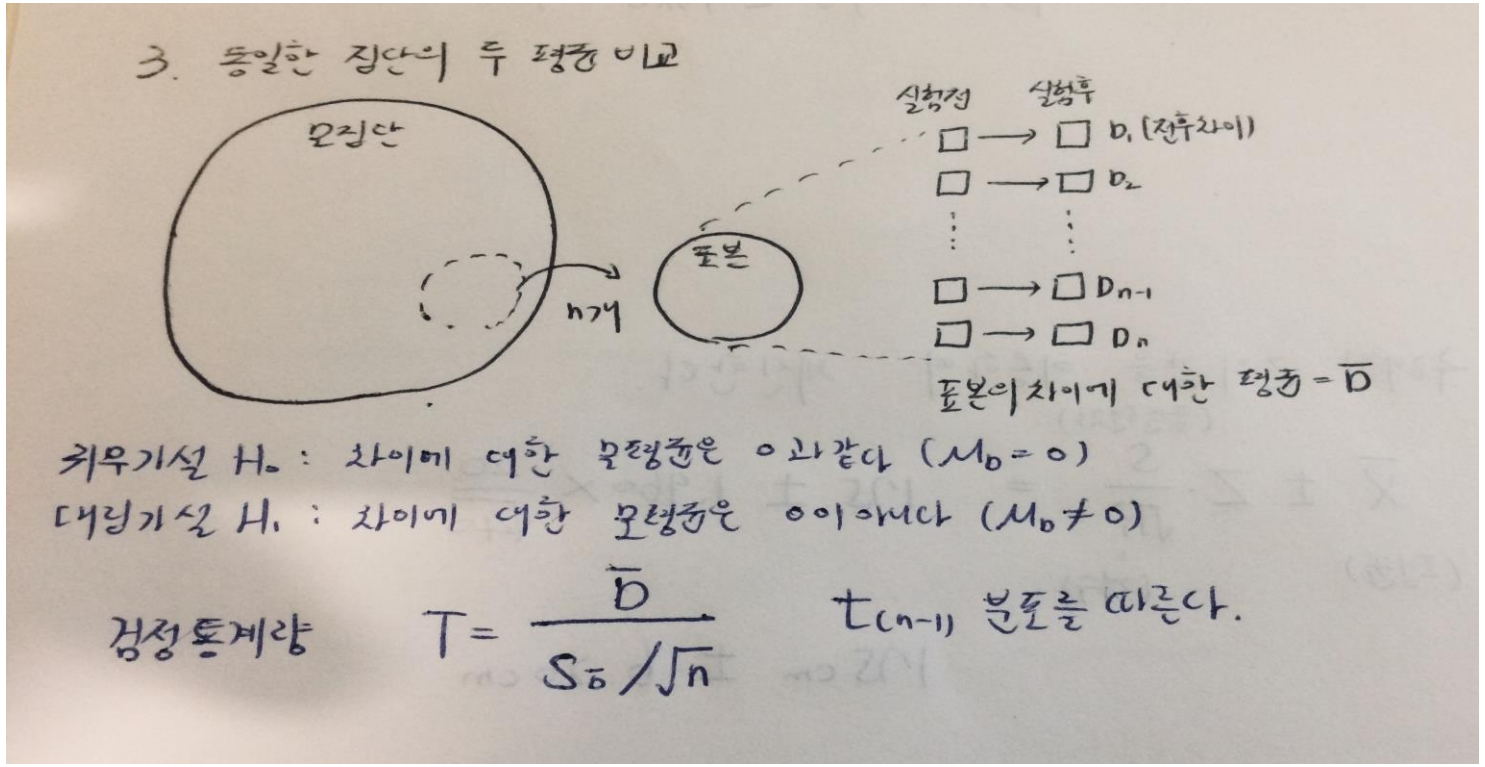
실제 효과가 있는 것을 통계적으로 효과가 있다고 보여 줄 수 있는 힘을 말한다.

연구자들은 제 1종 오류를 5%로 유지하면서 검정력을 최대화하는 통계 기법을 사용하고자 한다.

1-8 가설 검정(1)

▶ 동일한 집단의 두 평균 비교

일반적으로 연구의 효과성 검증을 위해 동일한 집단을 대상으로 실험 전후의 결과값을 비교하여 차이가 존재하는지를 검증한다.

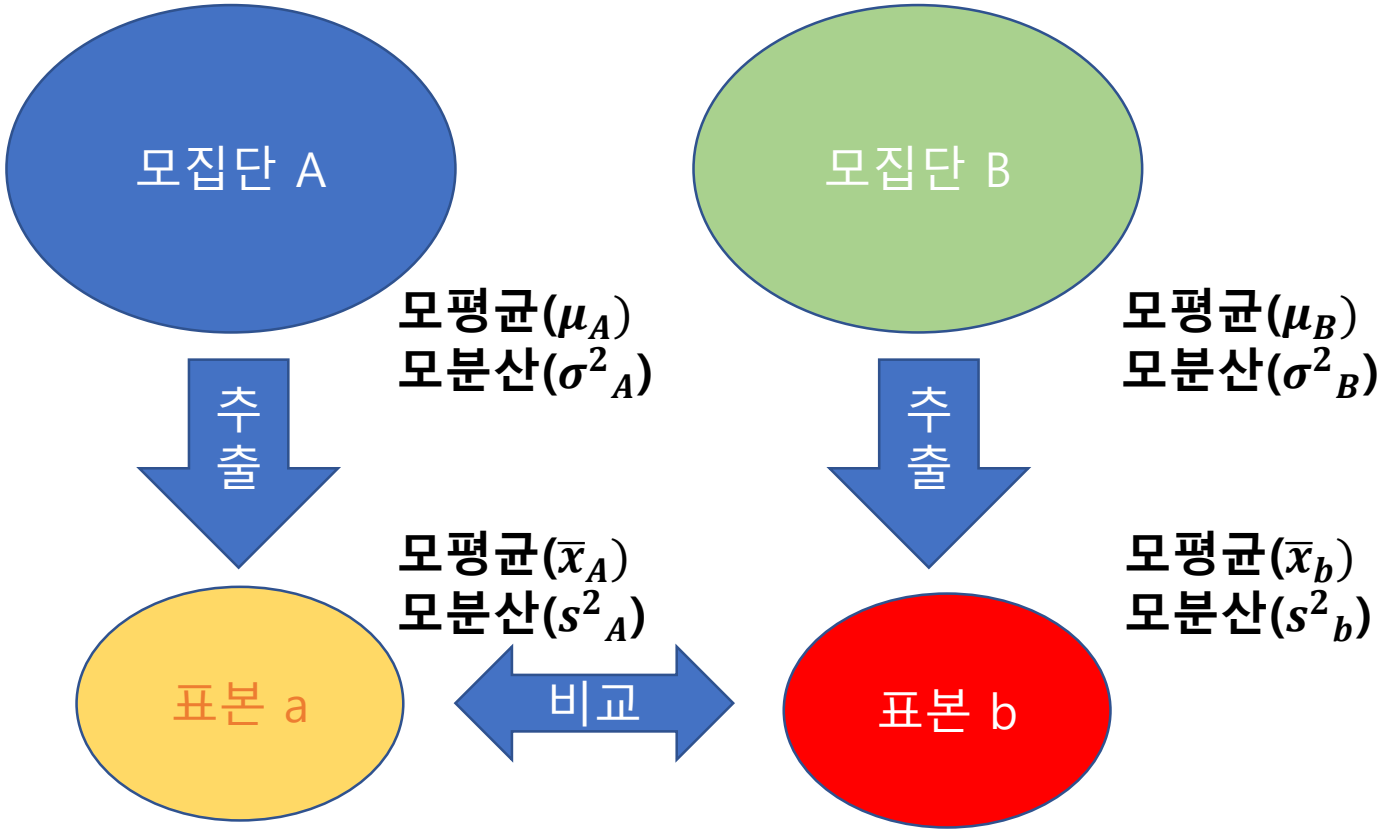


$\bar{D}$: 실험전후의 차의 평균
n : 관측치의 개수
$S_{\bar{D}}$: 표준편차

1-8 가설 검정(2)

▶ 두 모집단의 평균 차이에 대한 가설검정(독립표본)

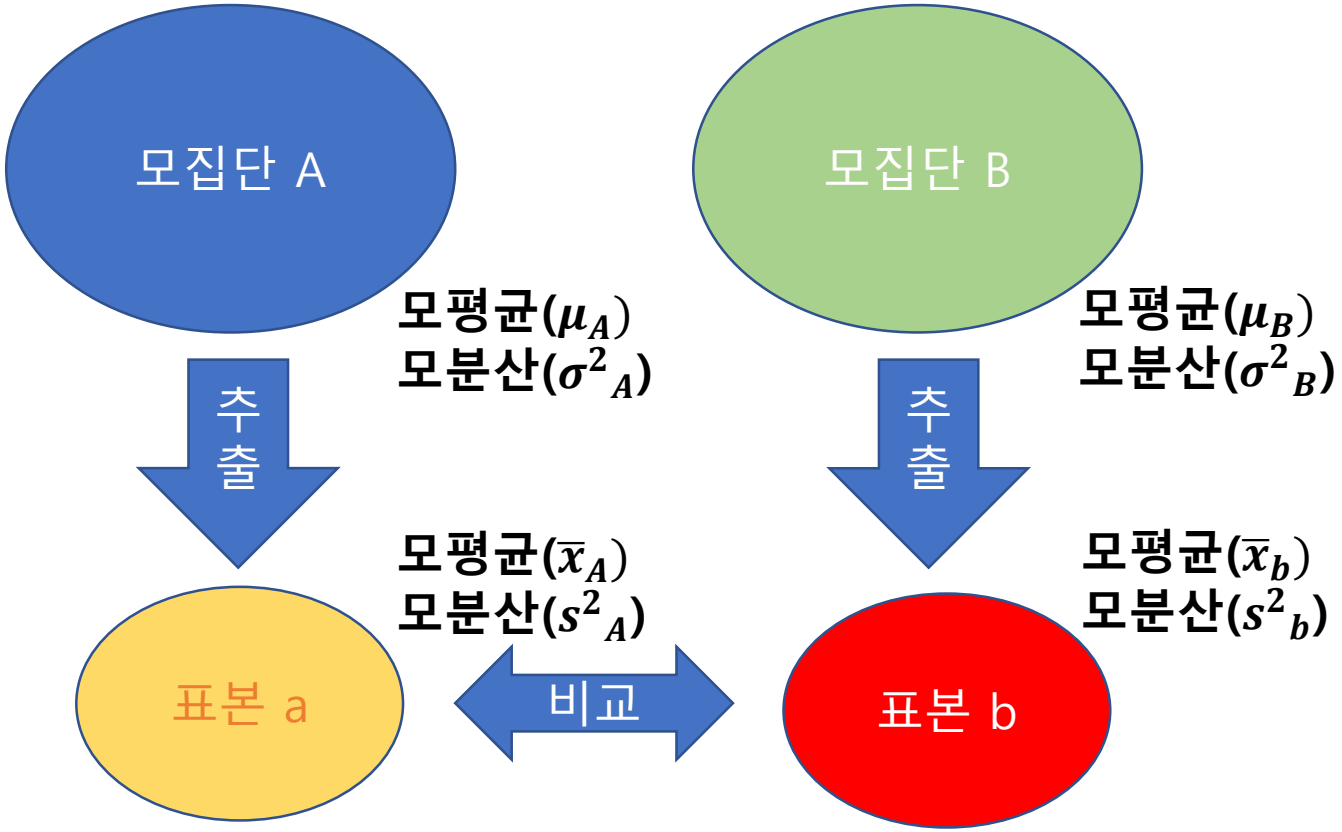
통계 대상의 대상이 되는 두 표본이 연관성이 서로 없는 독립 표본인 경우의 모집단



1-8 가설 검정(2)

▶ 두 모집단의 평균 차이에 대한 가설검정(독립표본)

통계 대상의 대상이 되는 두 표본이 연관성이 서로 없는 독립 표본인 경우의 모집단



독립 표본은 표본의 개수와 분산의 동일성 여부에 따라 네 가지 경우로 구분하여 신뢰구간과 검정통계량을 구할 수 있다.

- 표본의 개수가 충분하고, 모분산이 동일한 경우
- 표본의 개수가 충분하고, 모분산의 동일성을 모름
- 표본의 개수가 충분하지 않고, 모분산이 동일함.
- 표본의 개수가 충분하지 않고, 모분산의 동일성을 모름

1-9 분산 분석(Annalysis Of Variance:ANOVA)

▶ 분산분석(ANOVA)

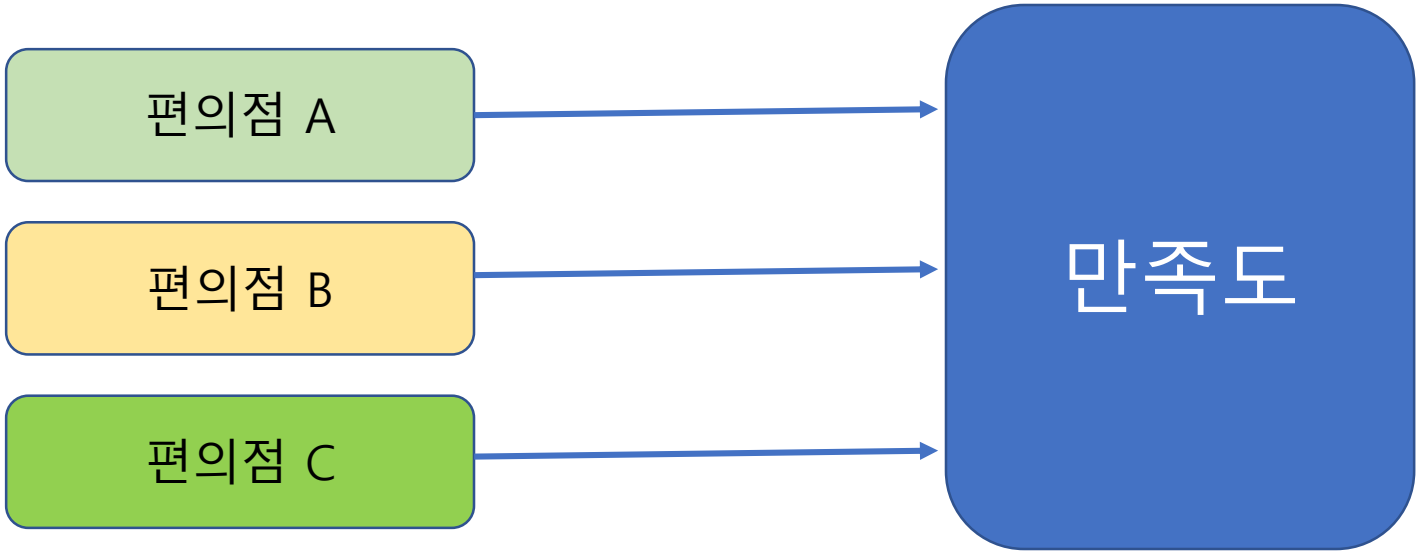
A. 3개 이상의 집단에 대한 평균 차이를 검증하는 방법

B. 3개이상의 집단에 평균 차이를 검정하기 위해 분산을 비교하는 분석 방법.

1-9 분산 분석(Annalysis Of Variance:ANOVA)

▶ 일원 분산분석(one-way ANOVA)

한 가지 요인을 기준으로 집단간의 차이를 조사하는 것.

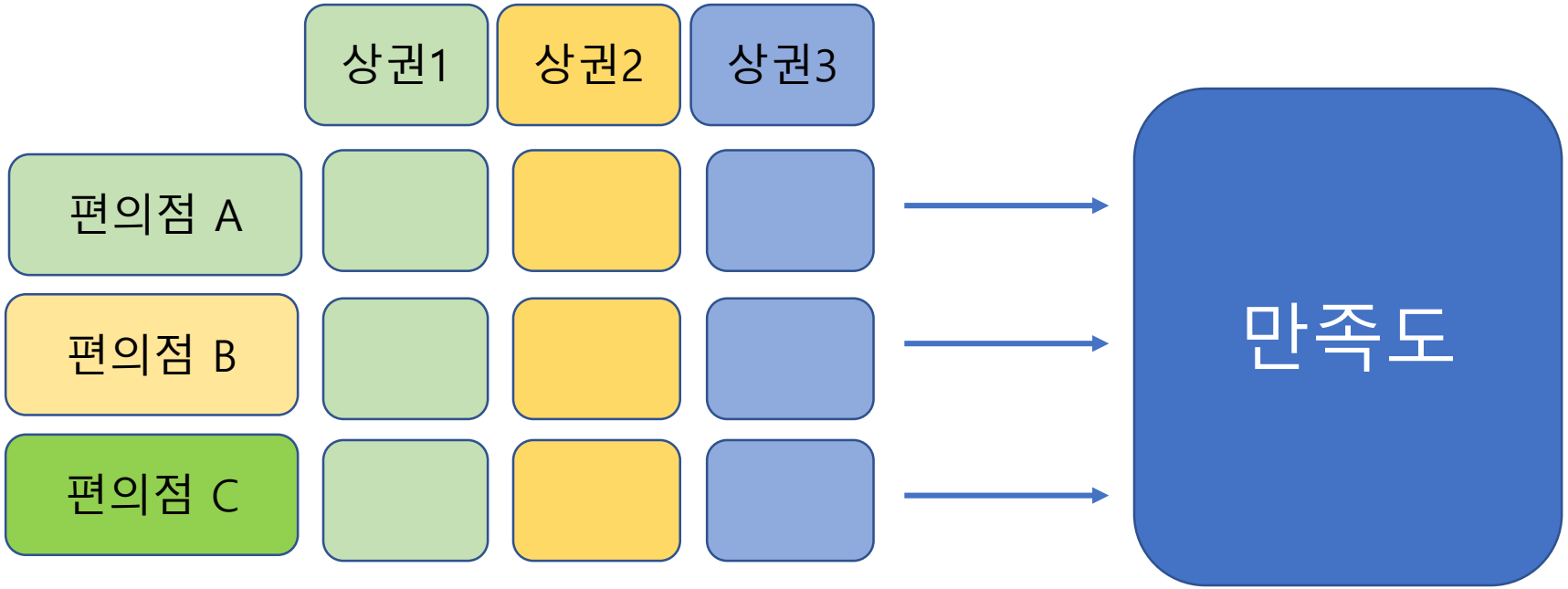


편의점의 종류를 기준으로 고객의 만족도를 조사하는 경우

1-9 분산 분석(Annalysis Of Variance:ANOVA)

▶ 이원 분산분석(two-way ANOVA)

두 가지 요인을 기준으로 집단간의 차이를 조사하는 것.



편의점의 종류와 위치를 기준으로 나누고, 편의점에 대한 고객의 만족도를 조사한다.

1-9 분산 분석(Annalysis Of Variance:ANOVA)

- **다원 분산분석(Multi-way ANOVA)**

  독립변수가 세 개 이상

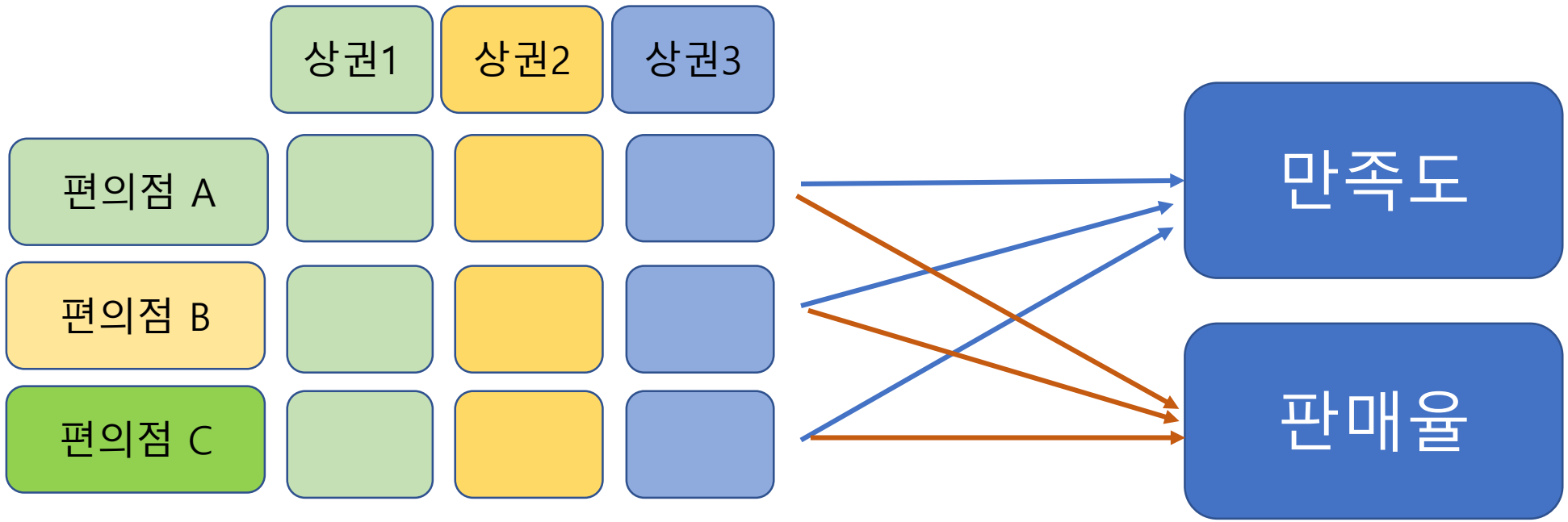
- **다변량 분산분석(MANOVA)**

  종속변수가 두 개 이상

1-9 분산 분석(Annalysis Of Variance:ANOVA)

▶ 다변량 분산분석(MANOVA)

종속변수가 두 개 이상



2-1 확률변수(random variable)

▶ 확률변수(random variable)

가. 확률 및 통계학에서 임의의 실험에 의해 얻어질 수 있는 모든 결과를 나타내는 변수

나. 확률변수는 무작위 사건에 숫자를 부여할 수 있다.

다. 확률 변수는 알려지지 않은 변수 또는 각 실험 결과에 할당하는 함수이다.

2-1 확률변수(random variable)

▶ 확률변수(random variable)

가. (예제) 하나의 주사위를 던질 때,

확률변수 X = “맨 위에 보여진 주사위 숫자“

X는 1,2,3,4,5,6일 수 있다.

나. (예제) 세 개의 동전을 던질 때,

확률변수 X = “앞면이 나올 동전의 수“

X는 0, 1, 2, 3일 수 있다.

2-1 확률변수(random variable)

▶ 확률변수(random variable)

다. (예제) 두 개의 주사위를 던질 때,

확률변수 X = "두 주사위의 값의 합"

		1st Die					
		1	2	3	4	5	6
2nd Die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

X는 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12이다.

(QA) 세개의 주사위를 던질 때, 세 주사위를 합계를 X로 했을 때, 확률변수 X는 무엇일까?

https://www.mathsisfun.com/data/random-variables.html 참조

REFERENCE

- 위키 백과(한, 영, 일)
- https://www.mathsisfun.com/data/random-variables.html
- 통계학 개론 – knou press
- 제대로 시작하는 기초 통계학 – 한빛 아카데미(노경섭 지음)
- 닥터 배의 술술 보건의학 통계 – 한나래(배정민 지음)

