Statistics II

for Machine Learning

Chap. 2: 추검정, ANOVA & 교차분석

Oh, Hyung Sool

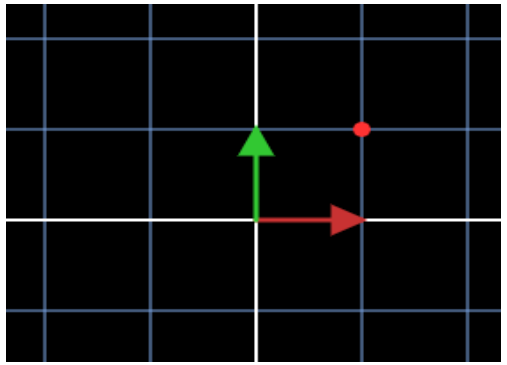
Matrix Product

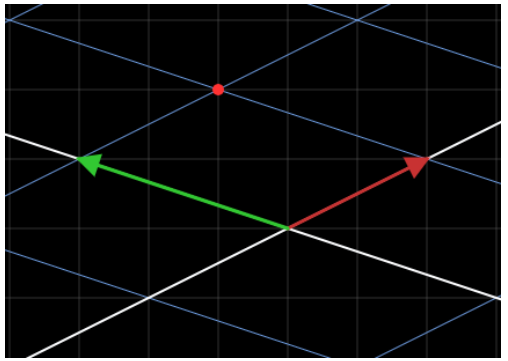
▶ 열벡터의 선형결합

$$\begin{cases} x+2y=3 \\ 3x+4y=5 \end{cases} \rightarrow \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \rightarrow x\begin{bmatrix} 1 \\ 3 \end{bmatrix} + y\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$$

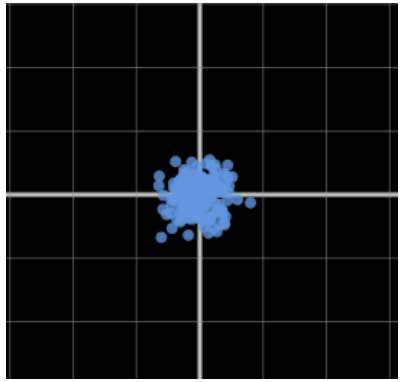
➢ ex) 다음의 행렬A를 이용하여 기저 벡터 $\vec{x}$ 를 변형

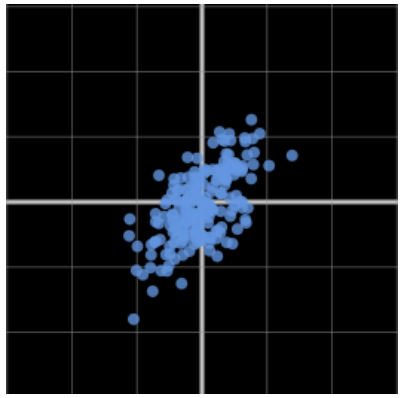
$$A = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix} \qquad \vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow A\vec{x} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$



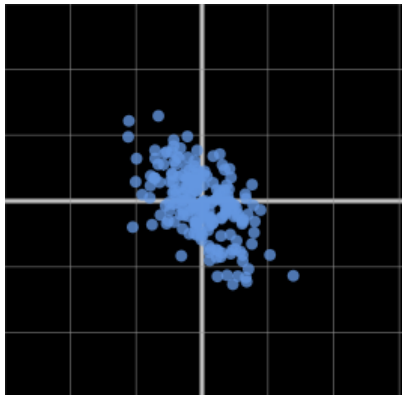


Matrix Product

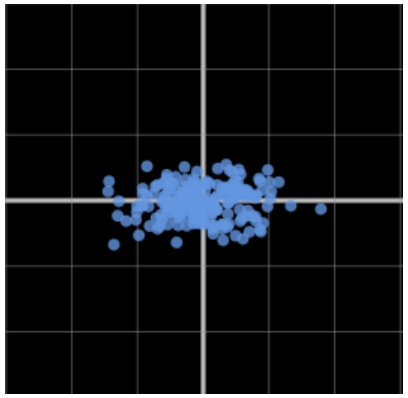




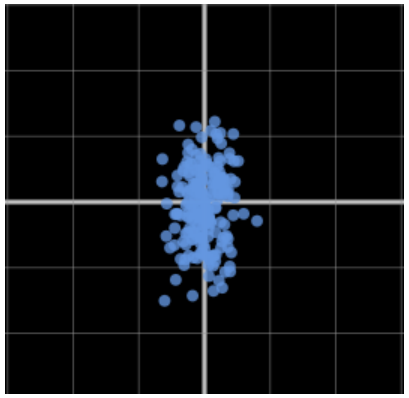
$$\begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 3 & -2 \\ -2 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

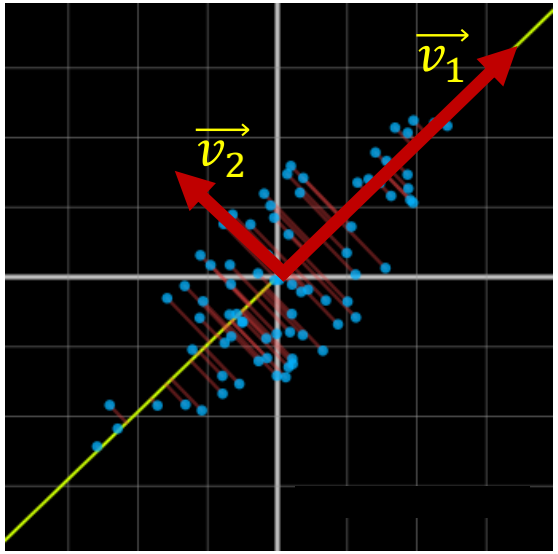
Matrix Product

An ***eigenvector*** **(**or **characteristic vector**) of a *linear transformation* is a nonzero vector that changes at most by a *scalar factor* when that linear transformation is applied to it.

$$\boldsymbol{AX} = \lambda \boldsymbol{X}$$

- $\boldsymbol{A}$: $matrix\ of\ the\ data, ie\ linear\ transformation$
- $\boldsymbol{X}$: $eigen\ vector\ for\ the\ linear\ transformation$
- $\boldsymbol{\lambda}$: $eigen\ value\ for\ the\ linear\ transformation$

- ***eigenvector***



$$A\vec{x} = \begin{bmatrix} 2 & -3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

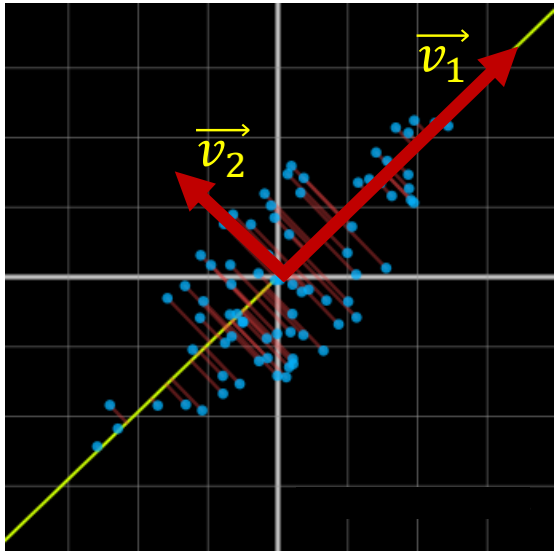
Eigen Vectors

$$C_1\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + C_2\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \ldots + C_k\begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

$$\begin{bmatrix} a_1 & \ldots & k_1 \\ \vdots & \ldots & \vdots \\ a_n & \ldots & k_n \end{bmatrix}\begin{bmatrix} C_1 \\ \vdots \\ C_k \end{bmatrix}$$

Eigenvector ??

$$\begin{bmatrix} a_1 & \ldots & k_1 \\ \vdots & \ldots & \vdots \\ a_n & \ldots & k_n \end{bmatrix}$$



Mean & Variance

• 우리는 회사에서의 업무처리 시에 **데이터**를 근거하여 보고하거나 의사결정을 한다.

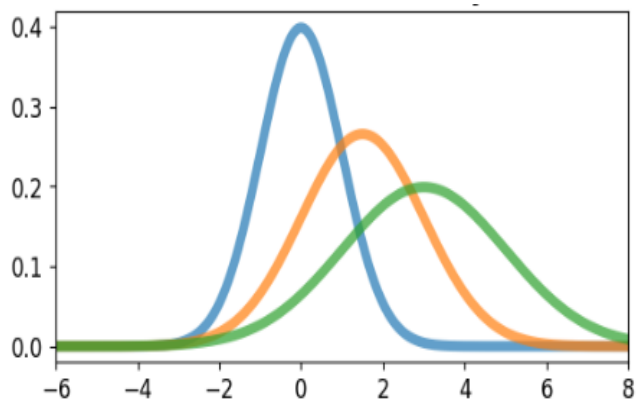
• **데이터**를 근거로 보고하거나 의사결정 하는 것이 곧 **통계적**으로 업무를 처리하는 것이다.

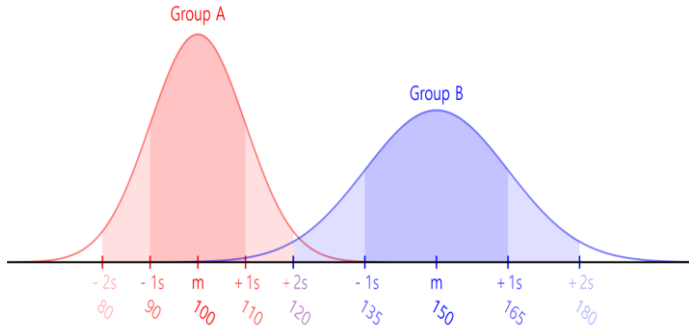
• 업무나 의사결정을 데이터에 근거하여 **통계적으로 처리할 때**, 반드시 사용하는 값이 **평균과 분산(편차)** 라는 통계량이다.

• (산술)평균의 의미와 이것이 주는 정보는 무엇인가?

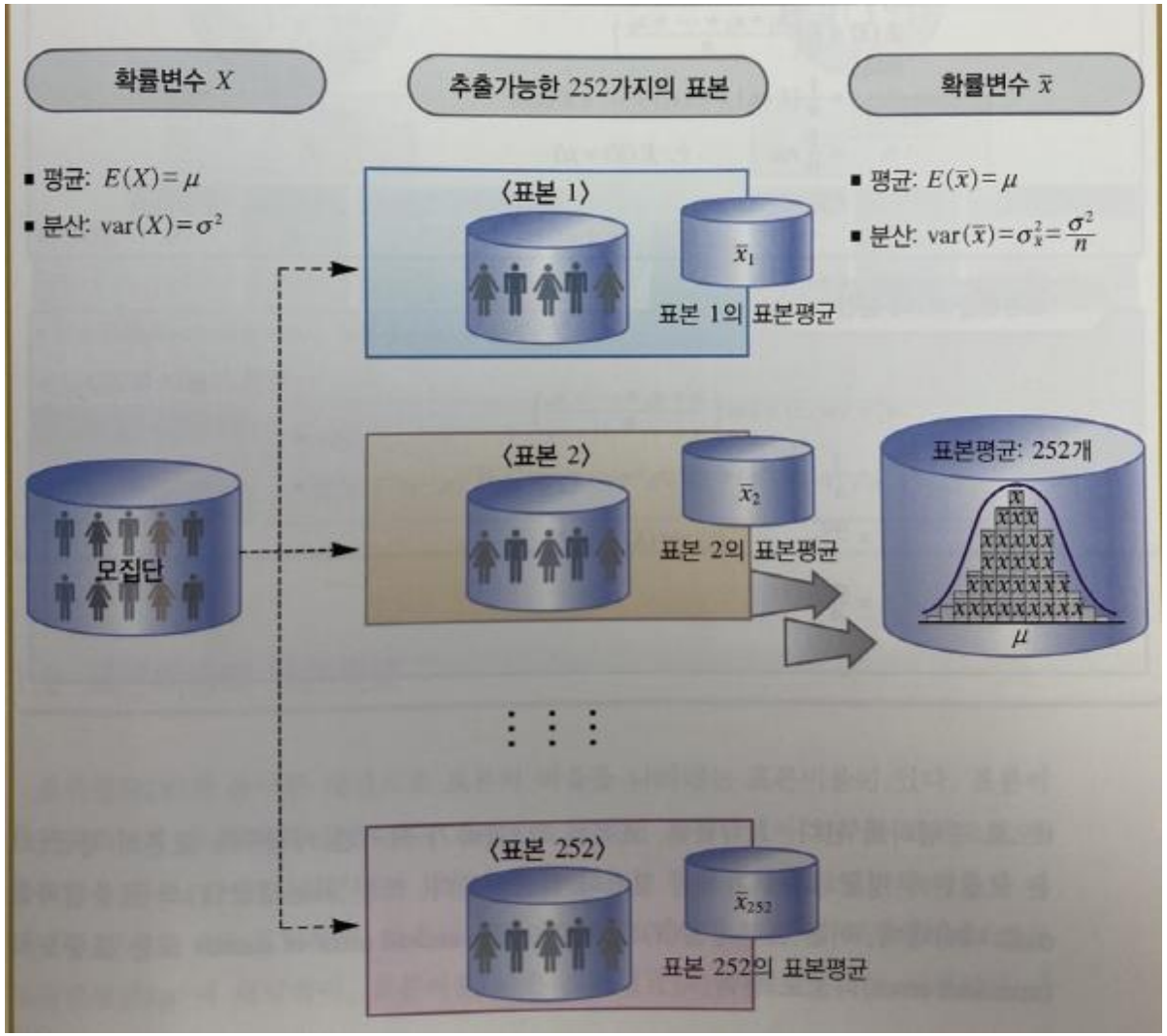
• 분산(편차)의 의미와 이것으로부터 얻는 정보는 무엇인가?

• 통계에서 즉 데이터를 처리하는데 있어서 **가장 중요한 것은 분산(편차)**인 이유는?





표분평균의 분포



● 표본평균: $\bar{x}=\frac{(x_1+x_2+\cdots+x_n)}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$

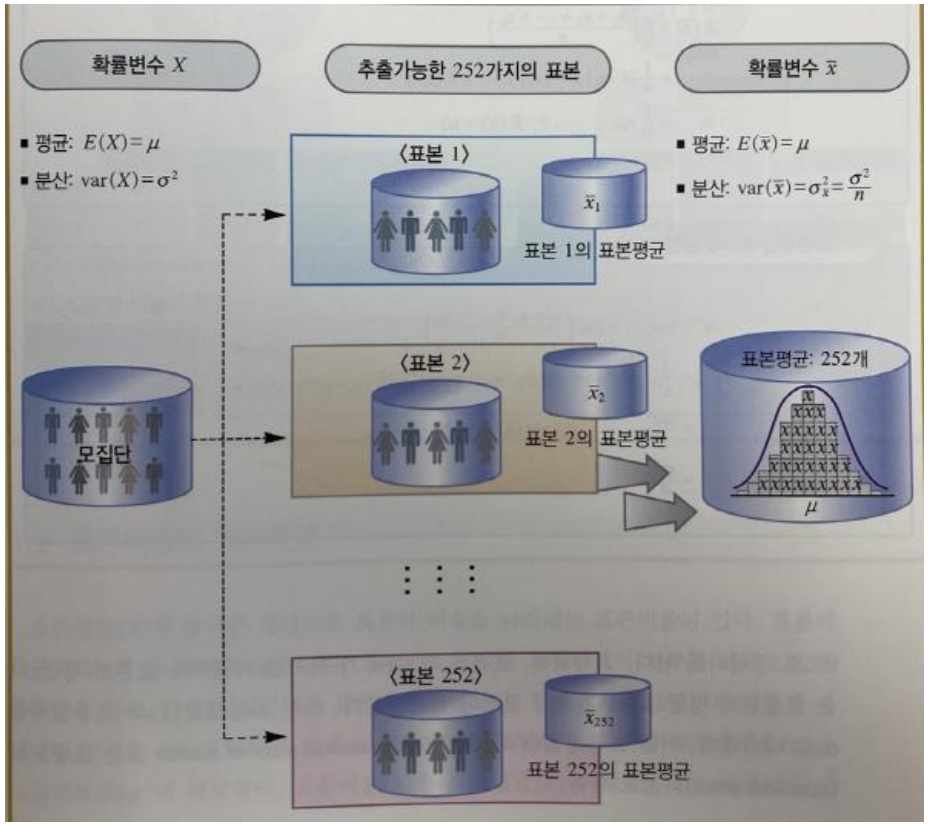
표본평균($\bar{x}$)의 기대값

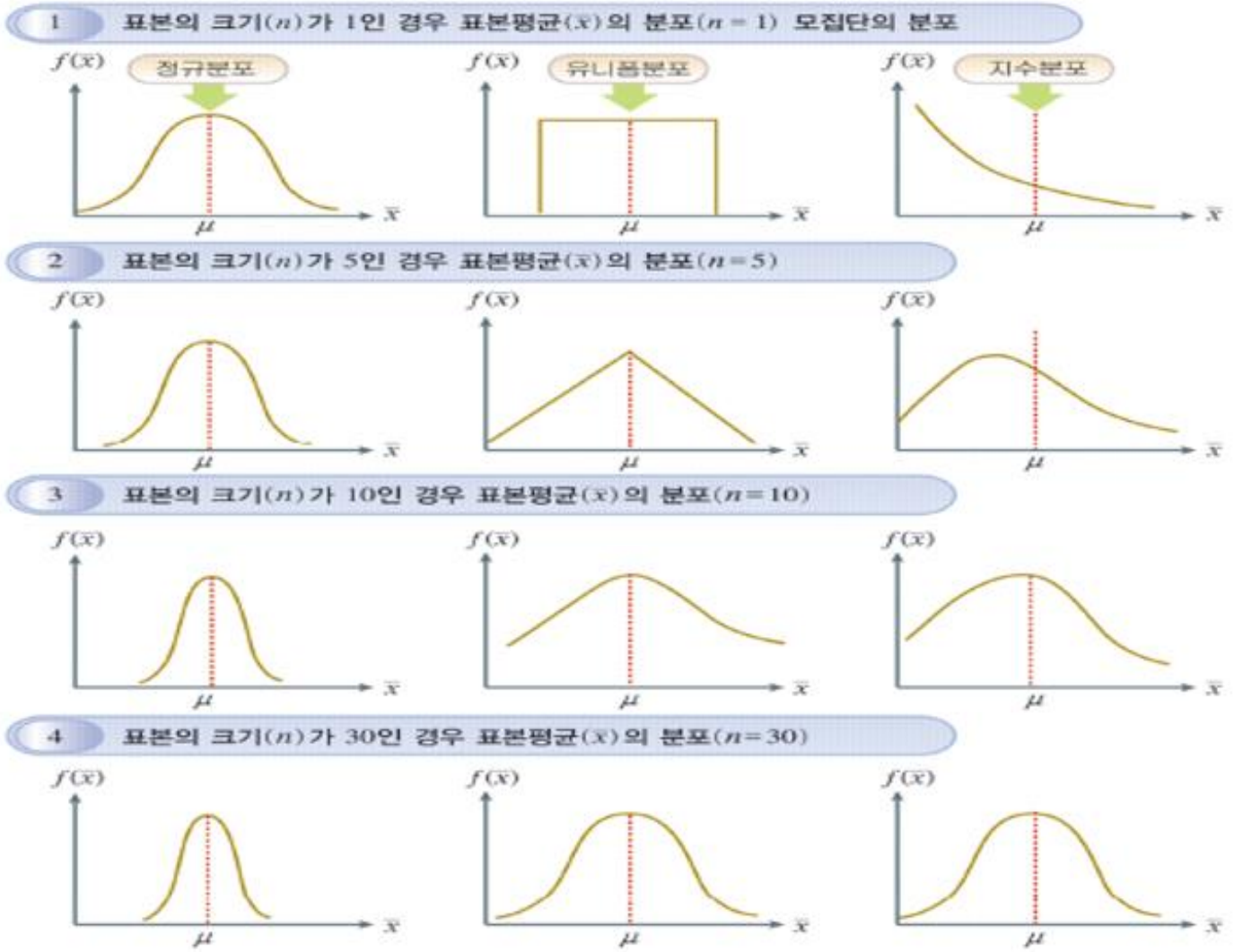
$$
\begin{aligned}
E(\bar{x}) &= E\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) \\
&= \frac{1}{n}[E(x_1)+E(x_2)+\cdots+E(x_n)] \\
&= \frac{1}{n}n\mu \qquad (\because E(X)=\mu) \\
&= \mu
\end{aligned}
$$

표본평균($\bar{x}$)의 분산

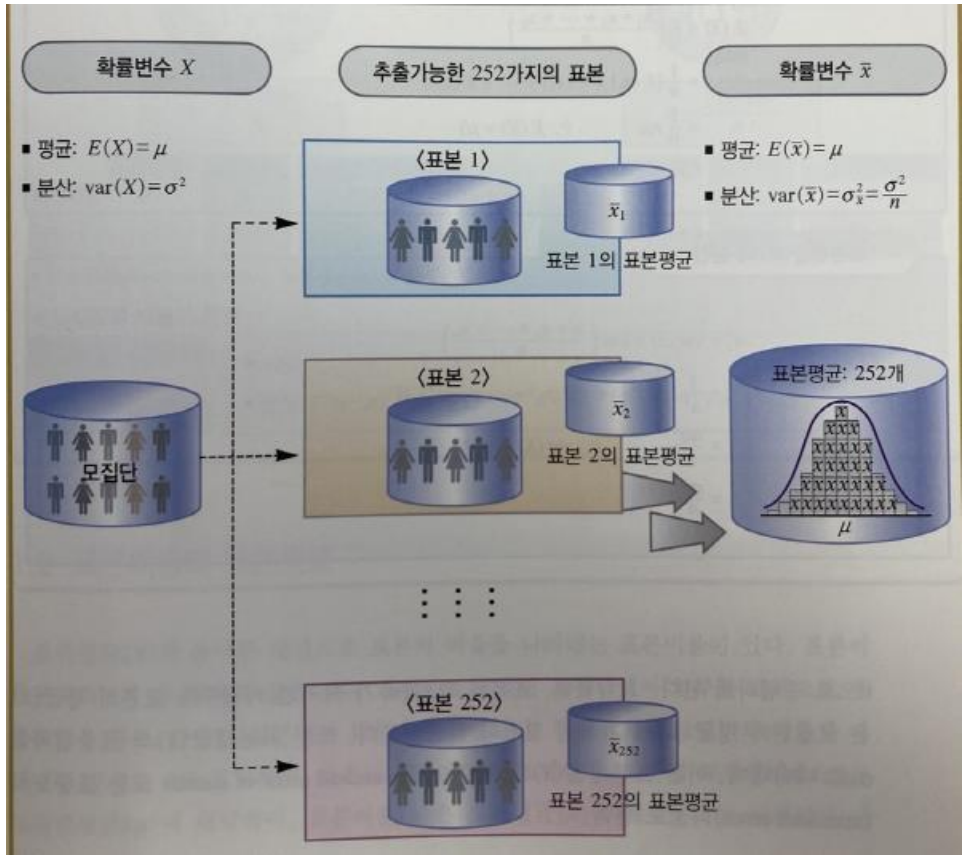
$$
\begin{aligned}
\sigma_{\bar{x}}^2 &= \mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) \\
&= \frac{1}{n^2}[\mathrm{var}(x_1)+\mathrm{var}(x_2)+\cdots+\mathrm{var}(x_n)] \\
&= \frac{n\sigma^2}{n^2} \qquad (\because \mathrm{var}(X)=\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

표분분포의 중심극한정리



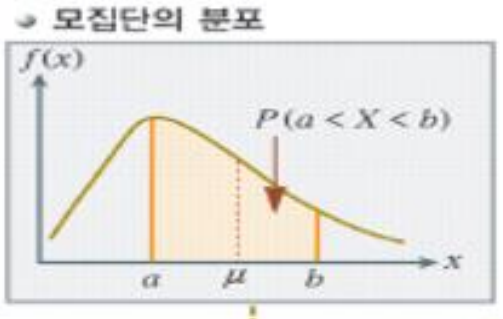


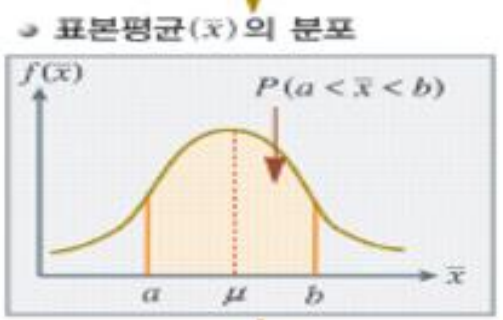
중심극한정리 & 표준화

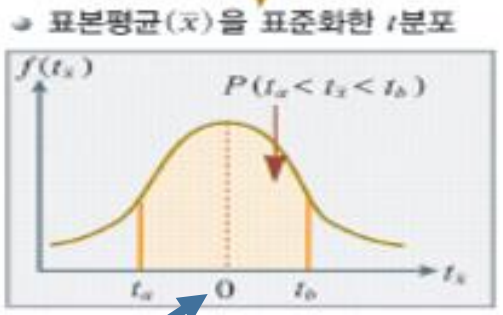


표준화

$$Z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$







중심극한정리

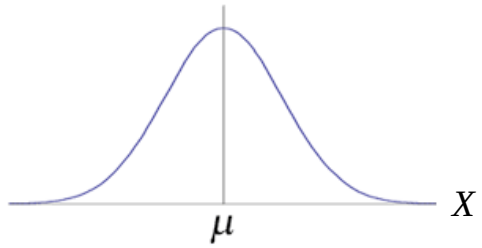
- 모집단이 어떠한 분포를 하는가에 관계없이 모집단으로부터 추출한 표본크기(n)가 커질수록 표본의 평균($\bar{x}$)은 정규분포에 준하는 분포를 하게 됨
- 일반적으로 표본크기(n)가 30개 이상이면, 약간의 예외적인 경우를 제외하고 모집단 분포와 관계없이 표본평균($\bar{x}$)의 분포는 거의 정규분포에 근접하는 분포를 함

자료의 표준화

- 표본평균($\bar{x}$)이 정규분포를 하여도 정규분포상의 특정 구간 내의 값을 가질 확률을 직접 구하는 것은 불가능함
- 따라서 표본평균($\bar{x}$)을 자유도가 ($n-1$)인 t분포상의 t값으로 표준화하여 확률값을 계산함
- 자유도가 ($n-1$)인 t분포상의 확률값은 미리 계산하여 부록으로 제시된 t분포의 확률표를 사용하면 쉽게 구할 수 있음

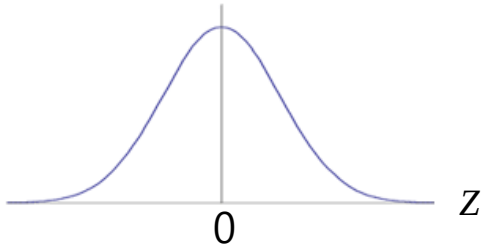
표준화의 필요성

ID	국어	영어
1	95	95
2	90	95
3	80	75
4	60	70
5	40	35
6	80	80
7	95	90
8	30	25
9	15	10
10	60	70
평균	64.5	64.5
분산	808	925
공분산	762	



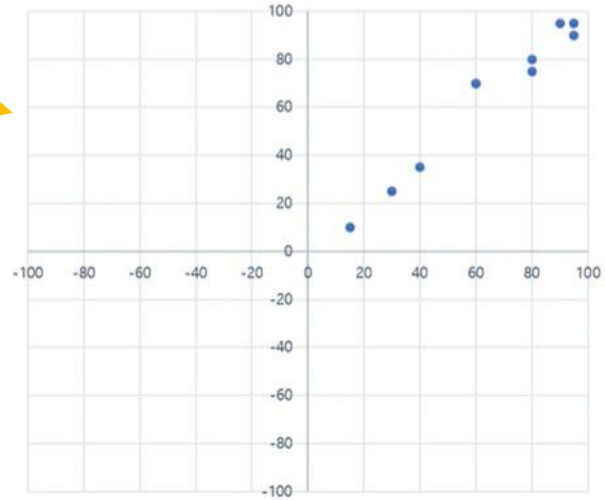
$$X \sim N(\mu, \sigma^2)$$

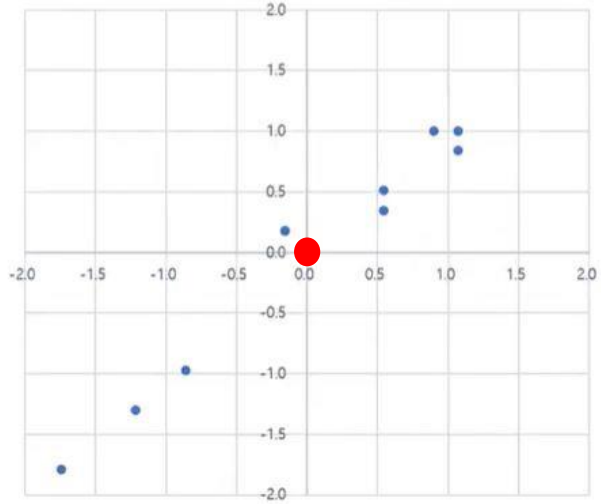
$$Z = \frac{x - \mu}{\sigma/\sqrt{n}}$$



$$Z \sim N(0, 1^2)$$

ID	국어	영어
1	1.1	1.0
2	0.9	1.0
3	0.5	0.3
4	-0.2	0.2
5	-0.9	-1.0
6	0.5	0.5
7	1.1	0.8
8	-1.2	-1.3
9	-1.7	-1.8
10	-0.2	0.2
평균	0	0
분산	1	1





ID	국어	영어
1	9.5	9.5
2	9.0	9.5
3	8.0	7.5
4	6.0	7.0
5	4.0	3.5
6	8.0	8.0
7	9.5	9.0
8	3.0	2.5
9	1.5	1.0
10	6.0	7.0
평균	6.45	6.45
분산	8.08	9.25
공분산	7.62	

표준화 결과의 특징

ID	국어	영어
1	95	95
2	90	95
3	80	75
4	60	70
5	40	35
6	80	80
7	95	90
8	30	25
9	15	10
10	60	70
평균	64.5	64.5
분산	808	925
공분산	762	

ID	국어	영어
1	9.5	9.5
2	9.0	9.5
3	8.0	7.5
4	6.0	7.0
5	4.0	3.5
6	8.0	8.0
7	9.5	9.0
8	3.0	2.5
9	1.5	1.0
10	6.0	7.0
평균	6.45	6.45
분산	8.08	9.25
공분산	7.62	

$$Z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

ID	국어	영어
1	1.1	1.0
2	0.9	1.0
3	0.5	0.3
4	-0.2	0.2
5	-0.9	-1.0
6	0.5	0.5
7	1.1	0.8
8	-1.2	-1.3
9	-1.7	-1.8
10	-0.2	0.2
평균	0	0
분산	1	1

공분산 : 0.97

- 변수들의 원래 데이터 값에 상관없이 평균:0, 분산:1 이 된다
- 이로 인해, 데이터의 크기(scale)로 인한 분산(변동의 크기)가 달라지는 영향을 배제할 수 있게 된다
- 총분산이 변수의 개수와 동일하게 된다

➢ **총분산 = 변수의 개수 = 고유값의 합**

추정(Estimation)

- **경제 분야**: 경제 지표 등을 통한 경제 계획 및 운영 활동 등에 추정을 사용

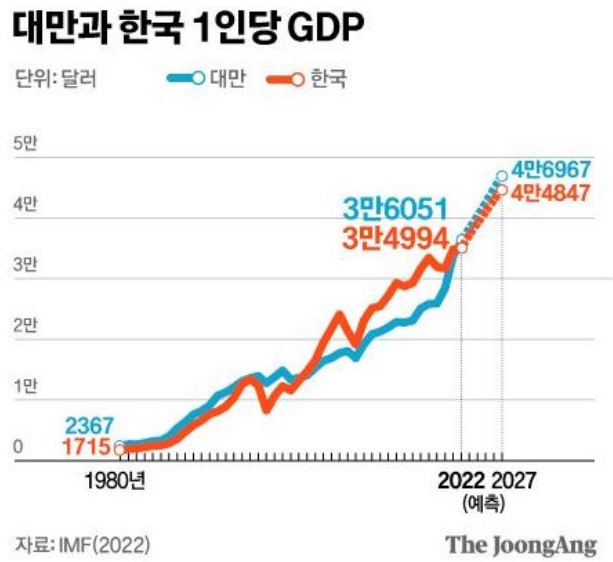
  예) 실업률, 물가상승률 등을 추정하기 위해

- **경영 분야**: 기업의 경영 성과, 고객 만족도 등을 통한 경영계획 및 전략 수립 등에 추정을 사용

  예) 매출액, 영업이익 등을 추정하기 위해

- **사회 분야**: 사회 현상, 사회 변화 등을 추정하는 데 사용

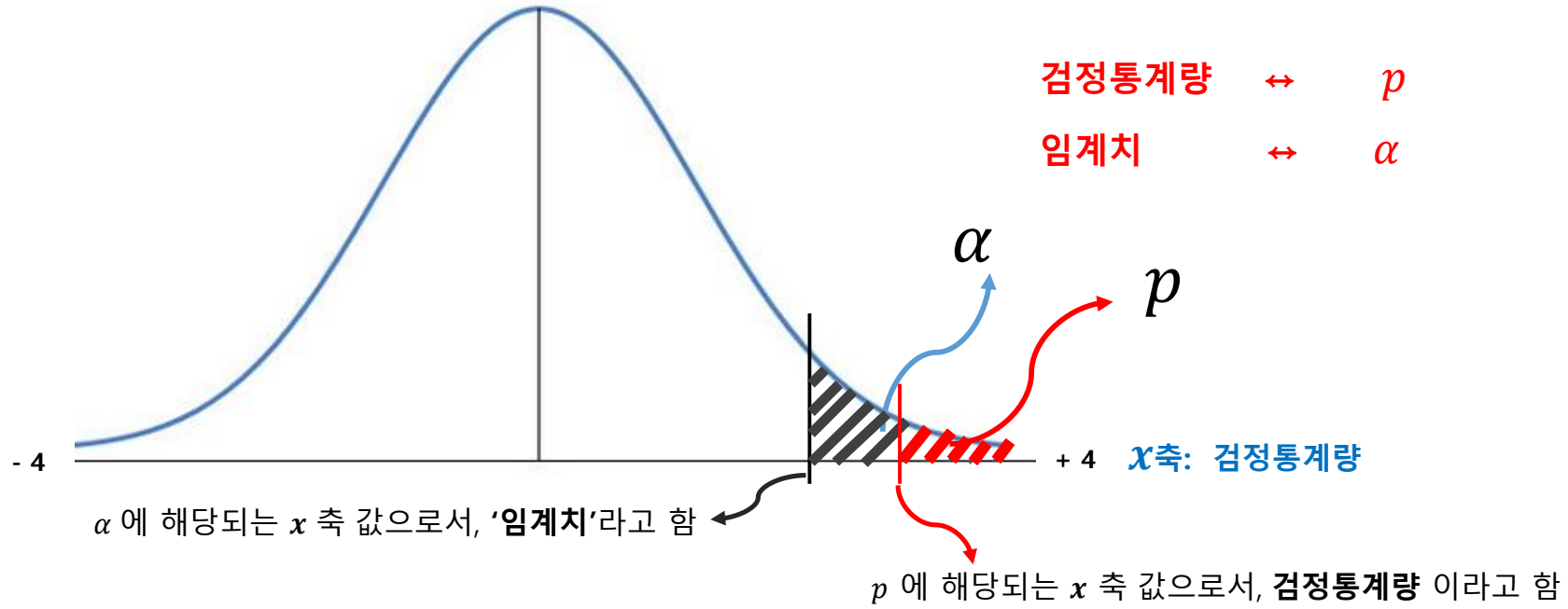
  예) 범죄율, 출산율 등을 추정하기 위해



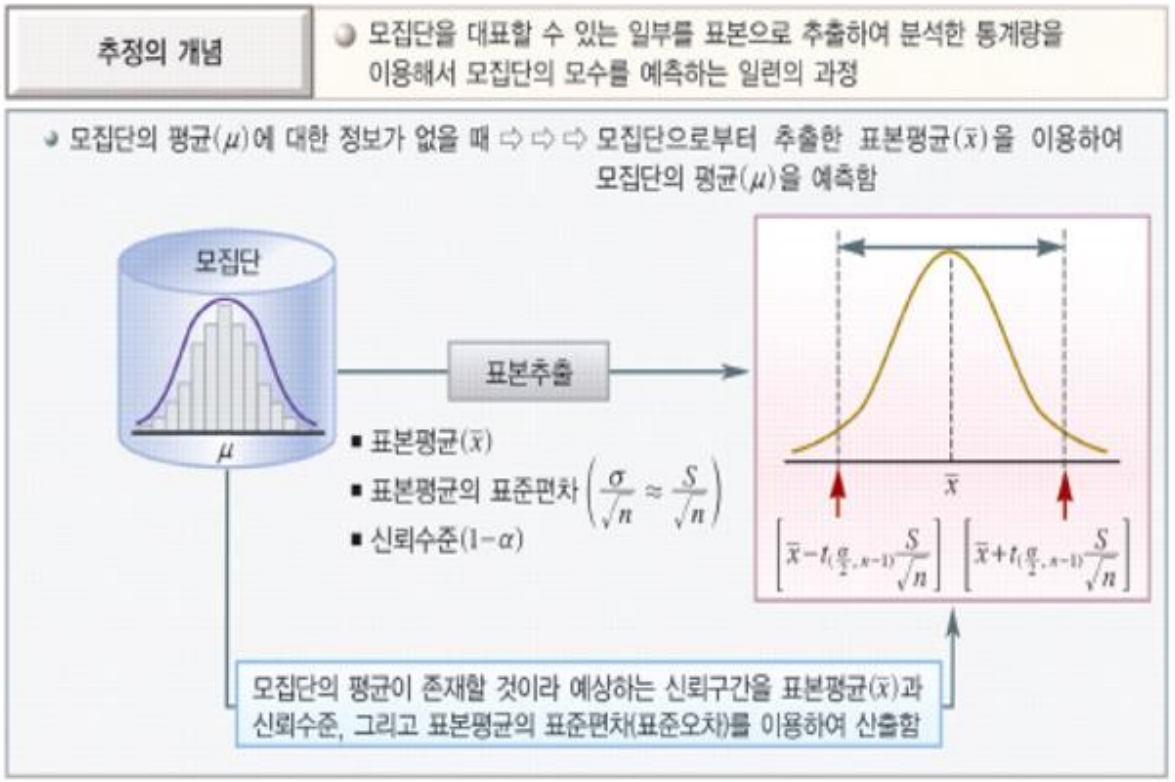
유의수준(α) vs. 유의확률(p)

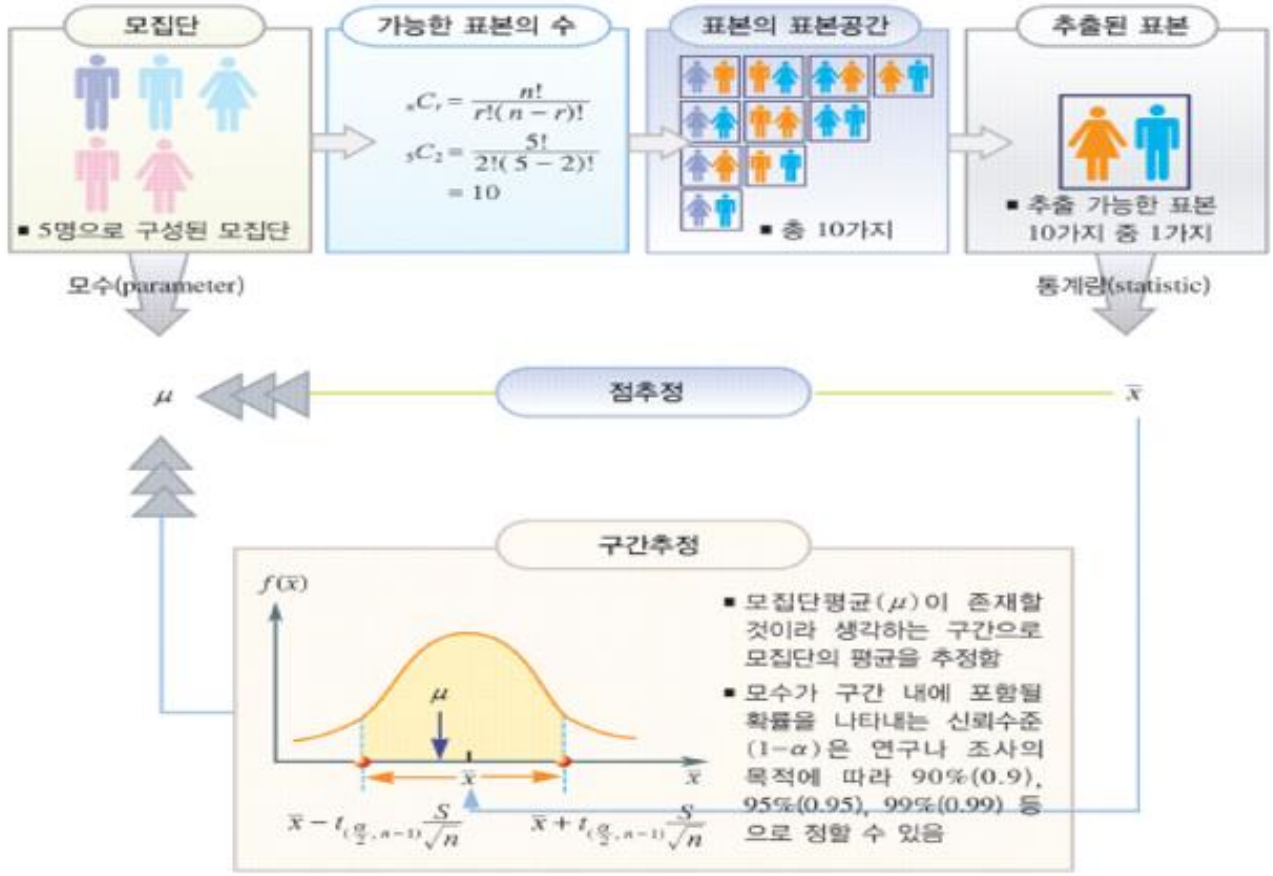
- 유의수준(α)이란?
  1) 통계량을 이용한 통계적 판정 결과가 틀릴(1종 과오) **가능성(probability)**
  2) 'Significance' 단어의 뜻 그대로, 다른 의미를 갖는다("**달라졌다**")라고 보는 기준

- 유의확률(p)이란?
  1) 귀무가설의 통계량 값으로 계산 한 **검정통계량 값**에 해당되는 **확률*(probability)***
  2) 'p 값이 α보다 작다($p<\alpha$)' 는 것은 달라졌다고 판정하는 기준 내에 해당된다는 의미



추정(Estimation)





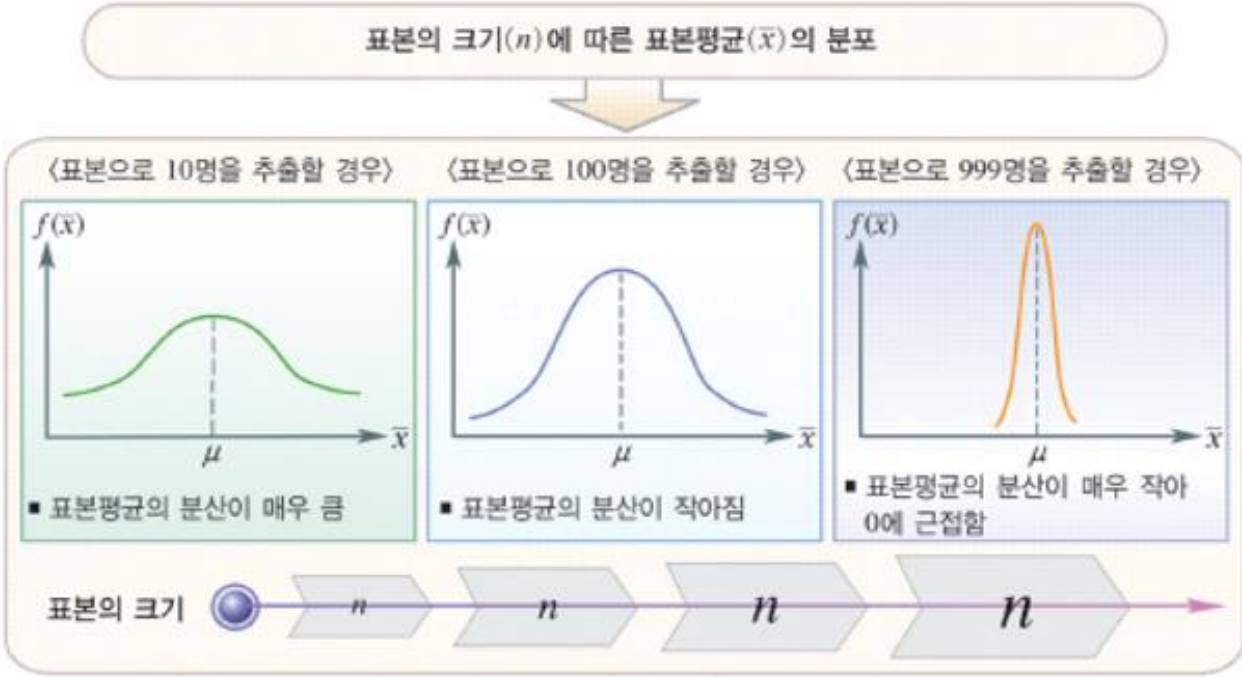
표준화 ➡ $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

추정(Estimation)

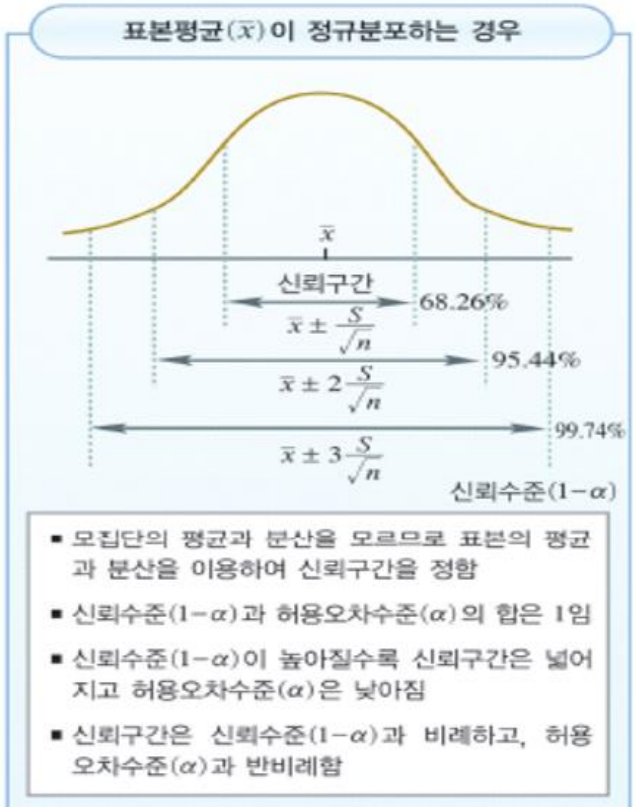
- 표본평균($\bar{x}$)의 분산:
  - $\text{var}(\bar{x}) = \text{var}\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right) = \frac{\text{var}(x_1) + \text{var}(x_2) + \cdots + \text{var}(x_n)}{n^2}$

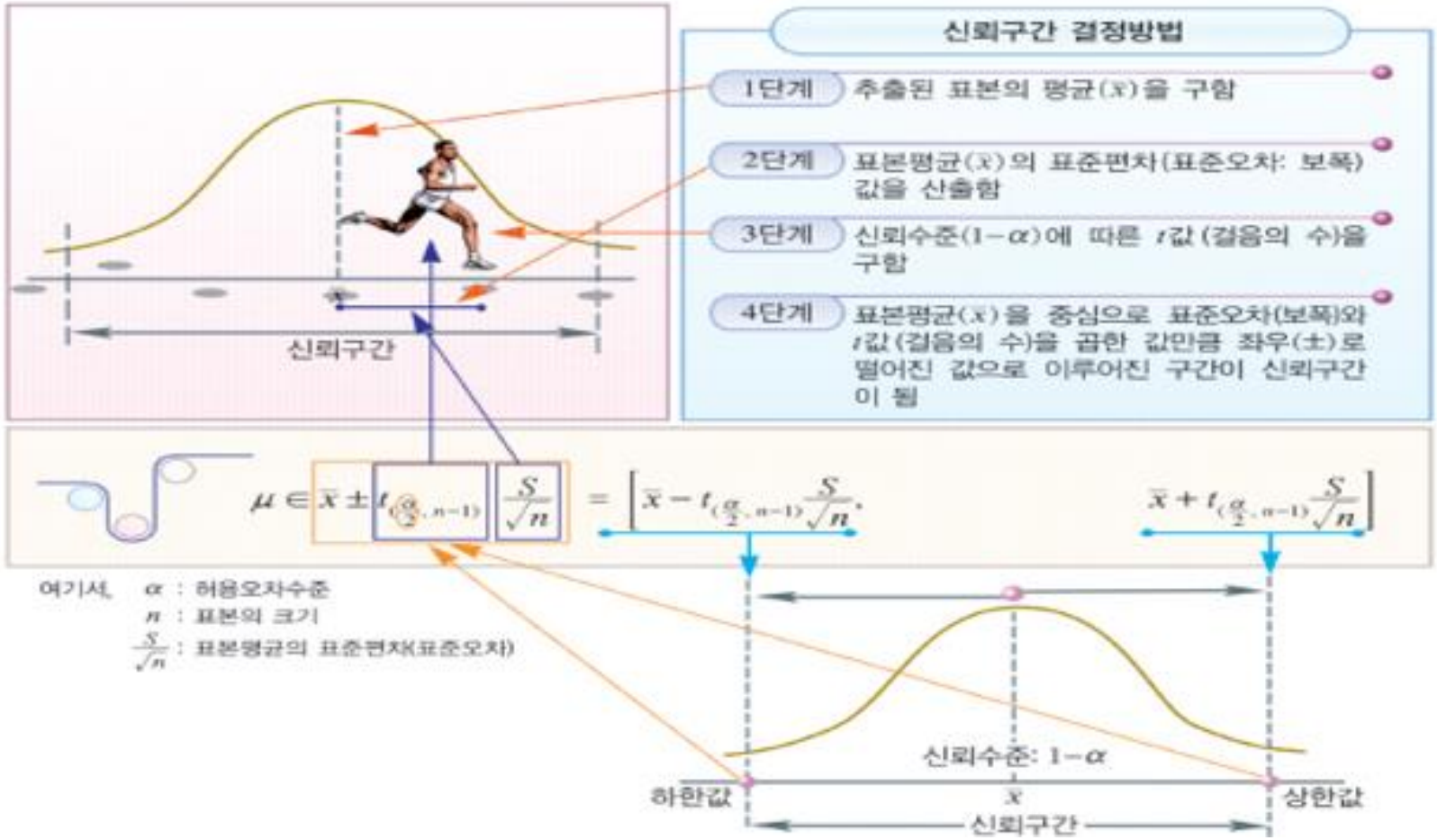
    $= \frac{n \times \text{var}(X)}{n^2} = \frac{\sigma^2}{n} \approx \frac{S^2}{n}$

- 표본평균($\bar{x}$)의 표준편차인 표준오차: $\frac{\sigma}{\sqrt{n}} \approx \frac{S}{\sqrt{n}}$



추정(Estimation): 신뢰구간





검정(Test)

데이터를 통해 전략(요인, 특성)의 영향력(효과)를 평가하는 데 활용

- **의학산업 분야**: 새로 개발하는 의약품의 효능과 안전성 등을 평가하는데 사용

- **경제 분야**: 경제정책 변경, 금리 조정, 무역정책 등의 효과를 평가하는데 사용

  예) 특정 정책이 실업율에 미치는 영향을 분석

- **경영 분야**: 마케팅 캠페인, 가격 전략, 제품 개발, 품질수준 향상 등의 효과를 평가하는데 사용

  예) 새로운 광고 캠페인이 매출에 미치는 영향을 분석

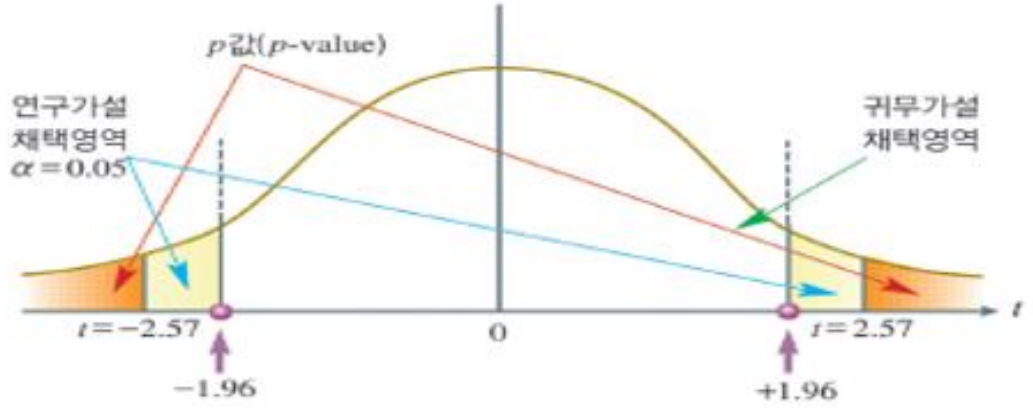
검정(Test)

가설검정

모수에 대한 새로운 가설이 옳다고 판단할 수 있는지를 표본통계량을 이용해서 평가하고 판단하는 과정

가설검정에서 사용되는 용어

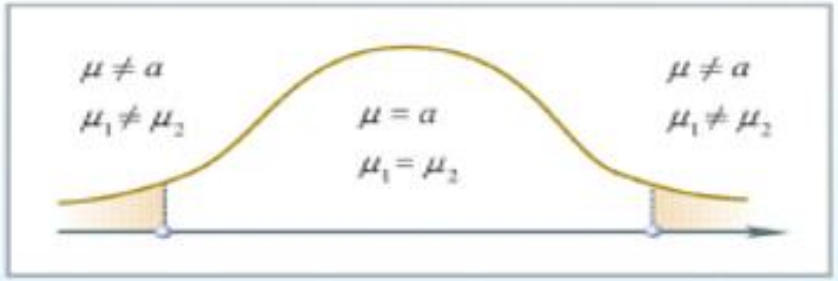
- **유의수준(α)**
  - 귀무가설이 옳다는 전제하에서 어느 정도 극단적인 표본통계량값이 나오면 귀무가설이 잘못되었다고 판단하여 귀무가설을 기각할 최대한의 확률
  - 귀무가설이 옳으나 귀무가설을 기각하고 연구가설을 채택할 1종 오류의 최대치
- **임계치**
  - 검정의 종류(양측, 단측)와 유의수준(α)을 고려해서 산출한 값으로 가설의 채택 여부를 결정짓는 경계값
- **검정통계량**
  - 표본으로부터 추출한 통계량이나 검정에 사용할 분포에 따라 그에 맞는 값으로 치환한 통계량
- **p값(p-value)**
  - 표본으로부터 얻은 통계량 혹은 이를 치환한 검정통계량의 절대값보다 더 큰 절대값을 또다른 표본으로부터 얻을 수 있는 확률(컴퓨터 패키지는 항상 양측검정을 기준으로 p값을 산출함)



검정(Test)

양측검정

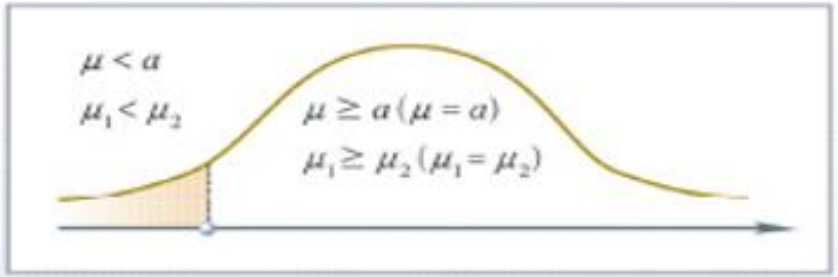
- 연구가설(H_1): $\mu \neq a$(단일 모집단)
  $\mu_1 \neq \mu_2$(두 모집단)
- 귀무가설(H_0): $\mu = a$(단일 모집단)
  $\mu_1 = \mu_2$(두 모집단)



단측검정

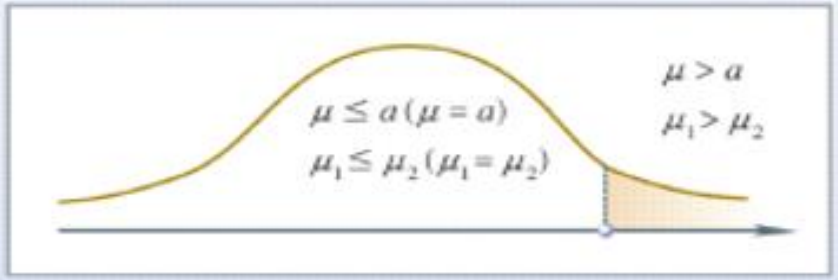
〈왼쪽꼬리검정〉

- 연구가설(H_1): $\mu < a$(단일 모집단)
  $\mu_1 < \mu_2$(두 모집단)
- 귀무가설(H_0): $\mu \geq a\,(\mu = a)$(단일 모집단)
  $\mu_1 \geq \mu_2\,(\mu_1 = \mu_2)$(두 모집단)

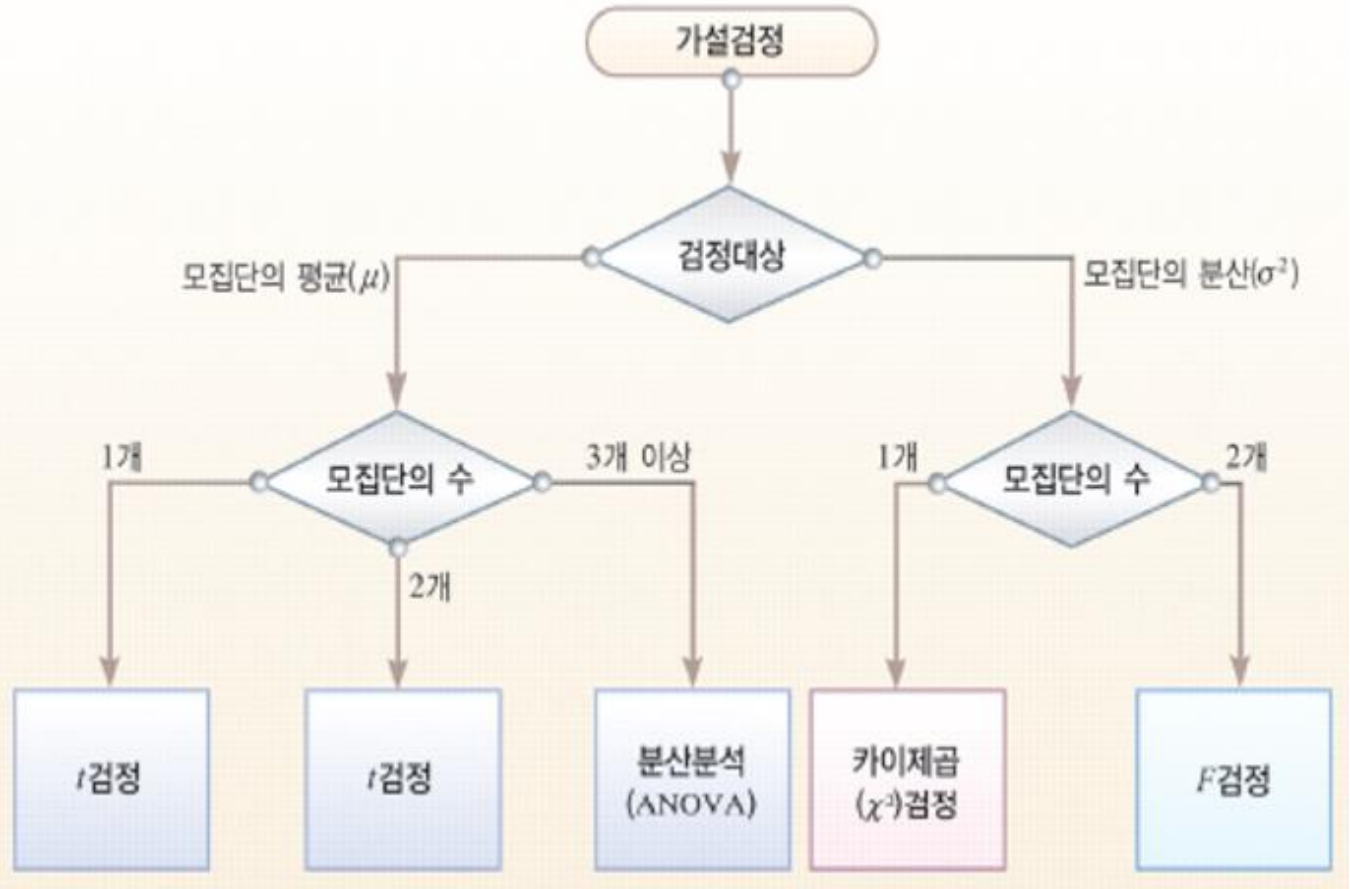


〈오른쪽꼬리검정〉

- 연구가설(H_1): $\mu > a$(단일 모집단)
  $\mu_1 > \mu_2$(두 모집단)
- 귀무가설(H_0): $\mu \leq a\,(\mu = a)$(단일 모집단)
  $\mu_1 \leq \mu_2\,(\mu_1 = \mu_2)$(두 모집단)



검정(Test)

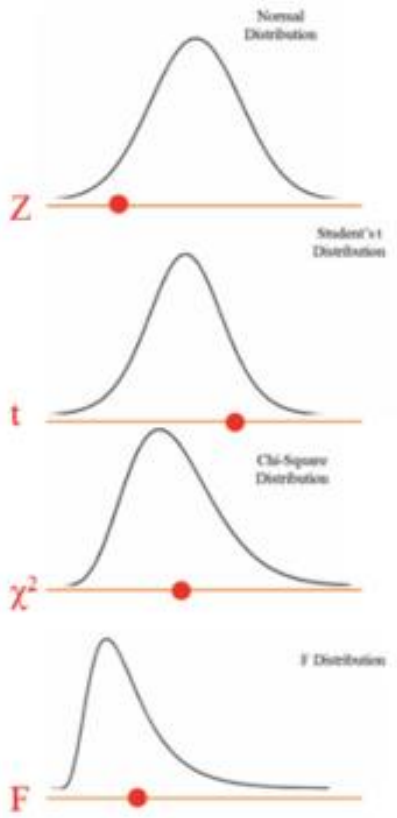


$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

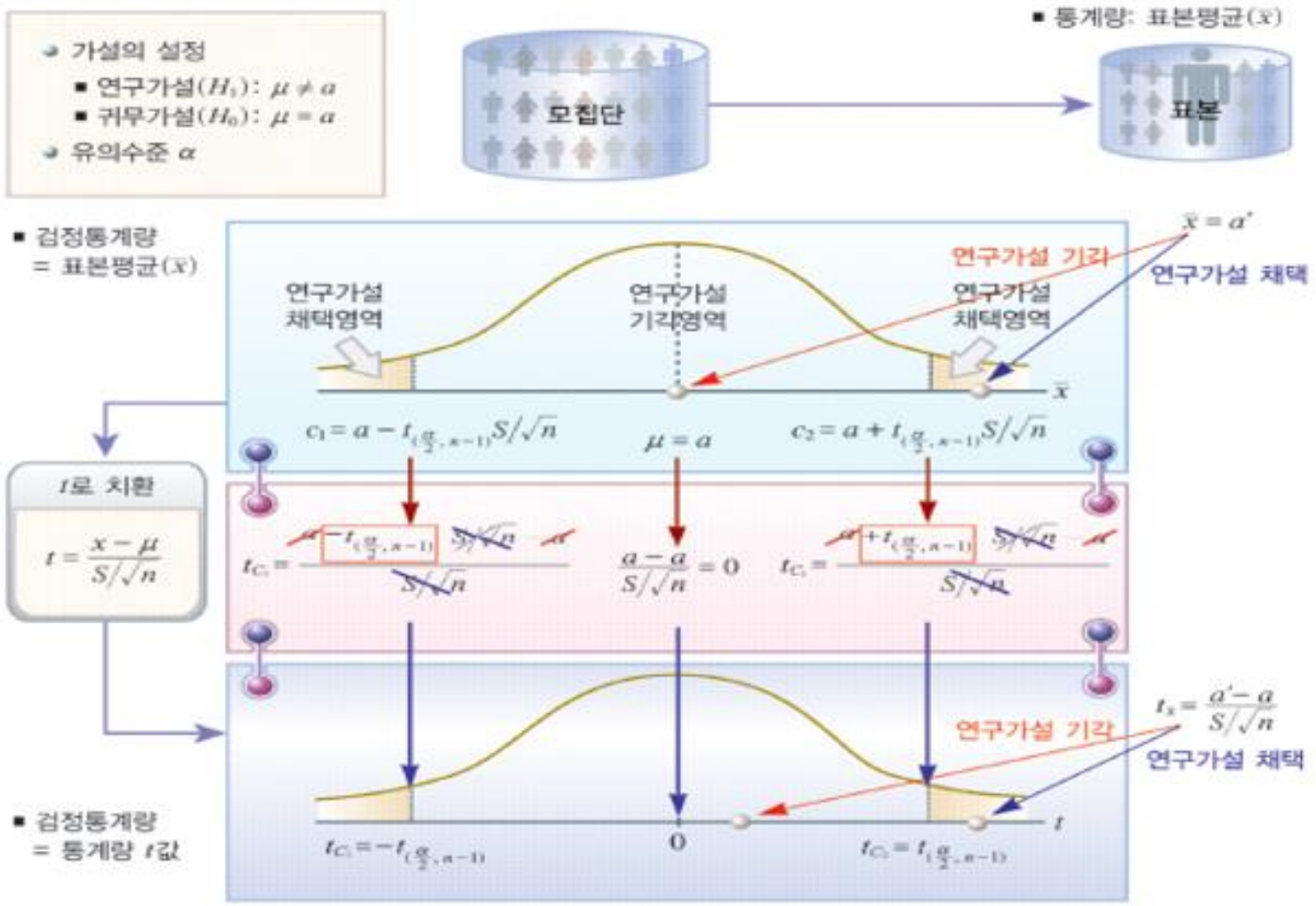
$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$F = \frac{s_1^2}{s_2^2}$$



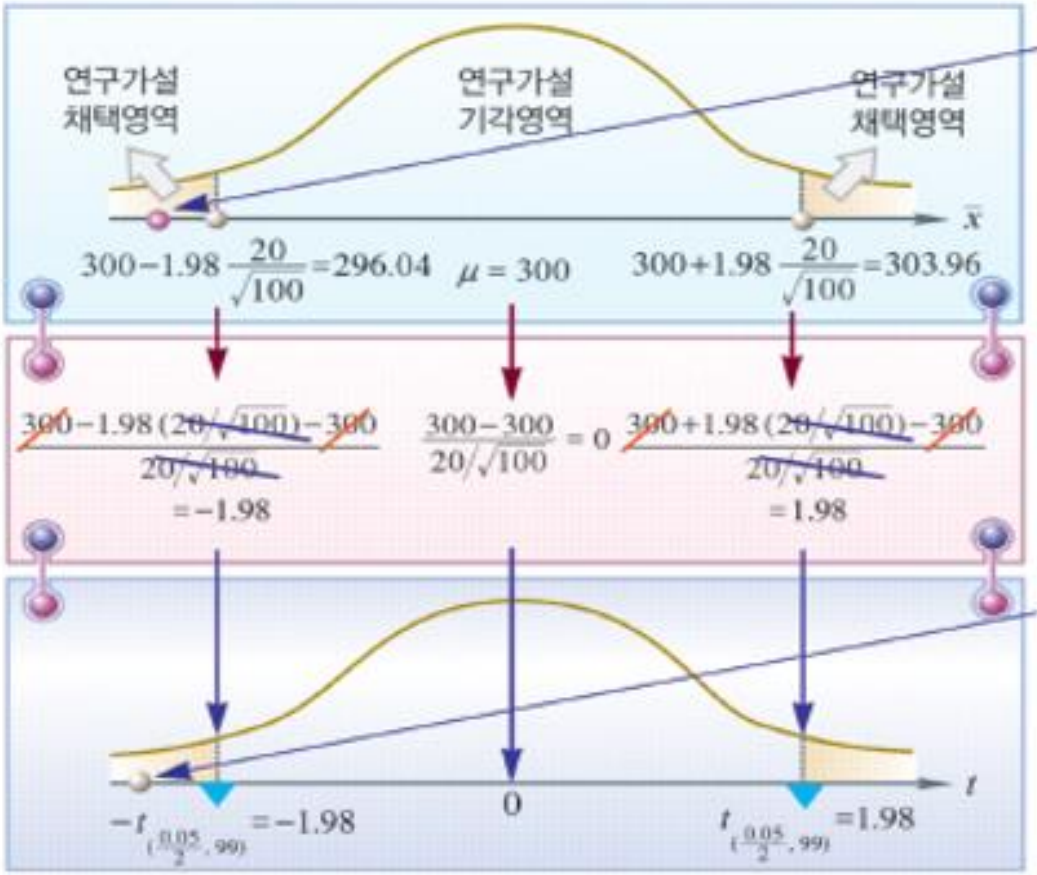
검정(Test): 평균


가설의 설정
연구가설(H_1): $\mu \neq a$
귀무가설(H_0): $\mu = a$
유의수준 α
모집단
표본
통계량: 표본평균($\bar{x}$)
검정통계량
= 표본평균($\bar{x}$)
연구가설 채택영역
연구가설 기각영역
연구가설 채택영역
연구가설 기각
연구가설 채택
$\bar{x} = a'$
$c_1 = a - t_{(\frac{\alpha}{2}, n-1)} S/\sqrt{n}$
$\mu = a$
$c_2 = a + t_{(\frac{\alpha}{2}, n-1)} S/\sqrt{n}$
t로 치환
$t = \frac{x - \mu}{S/\sqrt{n}}$
$\frac{a - a}{S/\sqrt{n}} = 0$
$t_x = \frac{a' - a}{S/\sqrt{n}}$
$t_C = -t_{(\frac{\alpha}{2}, n-1)}$
0
$t_C = t_{(\frac{\alpha}{2}, n-1)}$
검정통계량
= 통계량 t값

검정(Test): 평균

- 가설의 설정 — 연구가설(H_1): 통조림의 무게는 300g이 아니다($\mu \neq 300g$)
  귀무가설(H_0): 통조림의 무게는 300g이다($\mu = 300g$)
- 유의수준 — $\alpha = 0.05$

- 검정통계량 = 표본평균($\bar{x}$) = 294g
- 표본분산(S^2) = $400g^2$
- 표준오차 $\left(\frac{S}{\sqrt{n}}\right) = \frac{20}{\sqrt{100}} = 2$



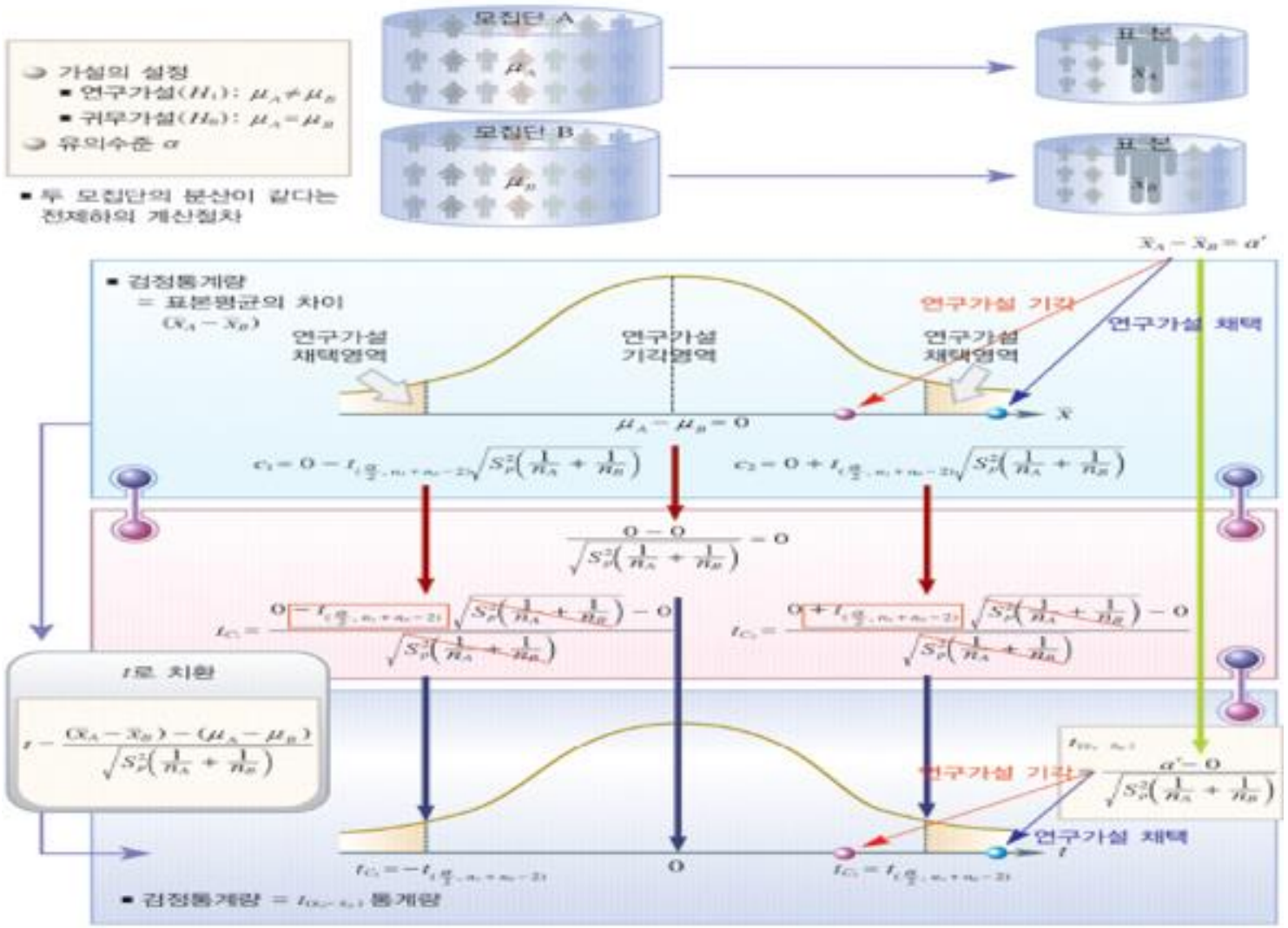
$\bar{x} = 294$
연구가설 채택

$t = \frac{294-300}{20/\sqrt{100}} = -3$

연구가설 채택

- 검정통계량 = 통계량 t값 $= \frac{294-300}{20/\sqrt{100}} = -3$

검정(Test): 평균의 차이



검정(Test): 평균의 차이

문 제

분당대학교에서는 두 개 반으로 나누어 통계학 강의가 실시되고 있다. 한 반은 주로 복학생들로 구성되어 있으며, 또 다른 반은 재학생들로 구성되어 있다. 담당 교수는 복학생들이 더 공부를 열심히 하여 시험 성적이 좋을 것이라 생각하고 있다. 실제로 그러한가를 검정하기 위하여 같은 내용의 시험을 보고 복학생 반과 재학생 반에서 각각 21명씩을 표본으로 추출하여 그 점수를 분석해 보았다.

	복학생반	재학생반
표본의 크기(n)	21	21
표본평균($\bar{x}$)	72.6	71.4
표준편차(S)	9.44	5.63

과연 교수님의 생각대로 복학생의 평균점수가 재학생의 평균점수보다 더 높다고 할 수 있는지에 대하여 5% 유의수준에서 검정해 보시오(단, 복학생과 재학생의 시험점수는 각각 정규분포하며 동일한 분산을 갖는다고 가정한다).

문제파악

- 연구가설(H_1): 복학생의 평균점수가 재학생의 평균점수보다 높다

  $H_1: \mu_{복학} > \mu_{재학}$ $(\mu_{복학} - \mu_{재학} > 0)$

- 귀무가설(H_0): 복학생의 평균점수가 재학생의 평균점수보다 낮거나 같다
  혹은 복학생의 평균점수는 재학생의 평균점수와 차이가 없다. 즉, 같다

  $H_0: \mu_{복학} \leq \mu_{재학}$ 혹은 $\mu_{복학} = \mu_{재학}$ $(\mu_{복학} - \mu_{재학} \leq 0$ 혹은 $\mu_{복학} - \mu_{재학} = 0)$

검정(Test): 평균의 차이

1단계 가설설정

- 연구가설(H_1): 통계학 수업을 듣는 복학생의 시험점수가 재학생보다 높다($\mu_{복학} > \mu_{재학}$)
- 귀무가설(H_0): 통계학 수업을 듣는 복학생의 시험점수가 재학생보다 낮거나 같다($\mu_{복학} \leq \mu_{재학}$)
  혹은 통계학 수업을 듣는 복학생과 재학생의 시험점수는 같다($\mu_{복학} = \mu_{재학}$)

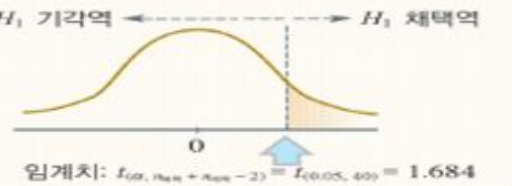
2단계 유의수준(α)과 검정의 종류에 따른 임계치 산출 및 연구가설의 채택영역 설정

- 설정된 가설을 보고 검정의 종류(양측검정, 왼쪽꼬리검정, 오른쪽꼬리검정)를 결정함

연구가설의 채택 및 기각영역 결정

오른쪽꼬리검정

- 연구가설(H_1): $\mu_{복학} > \mu_{재학}$
- 귀무가설(H_0): $\mu_{복학} \leq \mu_{재학}$ 혹은 $\mu_{복학} = \mu_{재학}$



- 표본통계량값이 클수록 연구가설(H_1)이 채택되는 오른쪽꼬리검정임
- 유의수준(α)이 0.05이며 자유도($n_{복학}+n_{재학}-2$)가 40인 t분포상의 임계치($t_{(0.05,\ 40)}$)는 1.684임
- 따라서 표본으로부터 구한 검정통계량 t값이 1.684보다 크면 연구가설을 채택하고 이보다 작으면 연구가설을 기각함

3단계 검정통계량(t값) 산출

- 귀무가설에서의 모평균간 차이가 0이므로 표본평균 차이를 표준오차로 나누어 검정통계량 t값을 구함

표본통계량

표본	복학생반	재학생반
n	21	21
$\bar{x}$	72.6	71.4
S	9.44	5.63

검정통계량(t)의 산출식

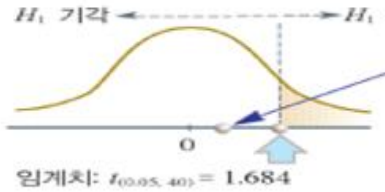
$$t = \frac{(\bar{x}_{복학} - \bar{x}_{재학}) - (\mu_{복학} - \mu_{재학})}{\sqrt{S_p^2\left(\frac{1}{n_{복학}} + \frac{1}{n_{재학}}\right)}}$$

검정통계량(t값)

$$t = \frac{(72.6 - 71.4) - 0}{\sqrt{60.41\left(\frac{1}{21} + \frac{1}{21}\right)}} = 0.488$$

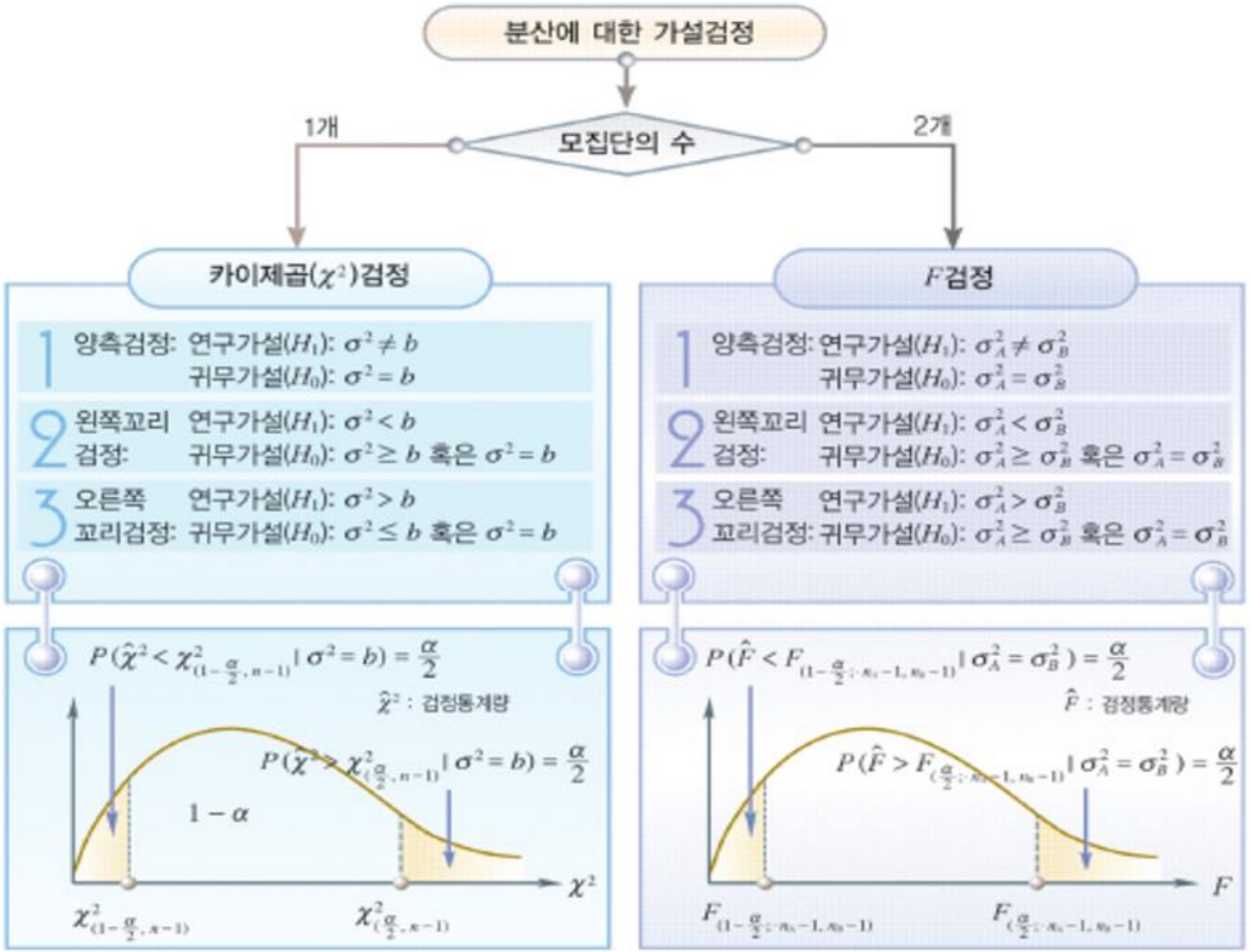
$$S_p^2 = \frac{(21-1)(9.44)^2 + (21-1)(5.63)^2}{21 + 21 - 2} = 60.41$$

4단계 임계치와 검정통계량을 비교하여 가설의 채택 여부를 결정함



- 임계치와 검정통계량을 비교한 결과, 검정통계량 t값(0.488)이 임계치(1.684)보다 작기 때문에 연구가설을 기각하고 귀무가설을 채택함
- 따라서 유의수준(α) 0.05에서 검정해본 결과 통계학 수업을 듣는 복학생반의 평균점수가 재학생반의 평균점수보다 통계적으로 유의할 정도로 높다고 말할 수 없음

검정(Test): 분산

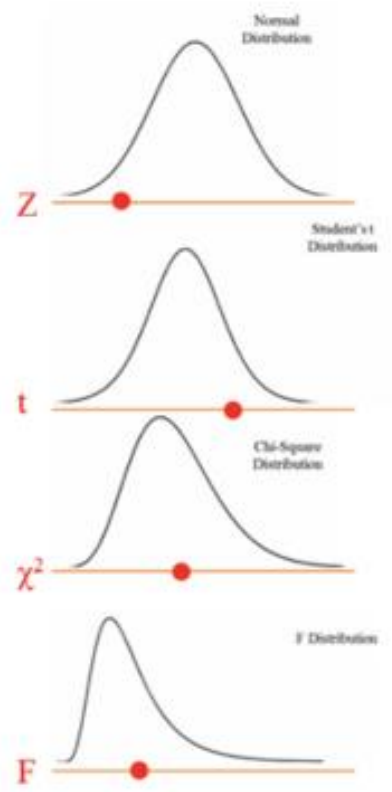


$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

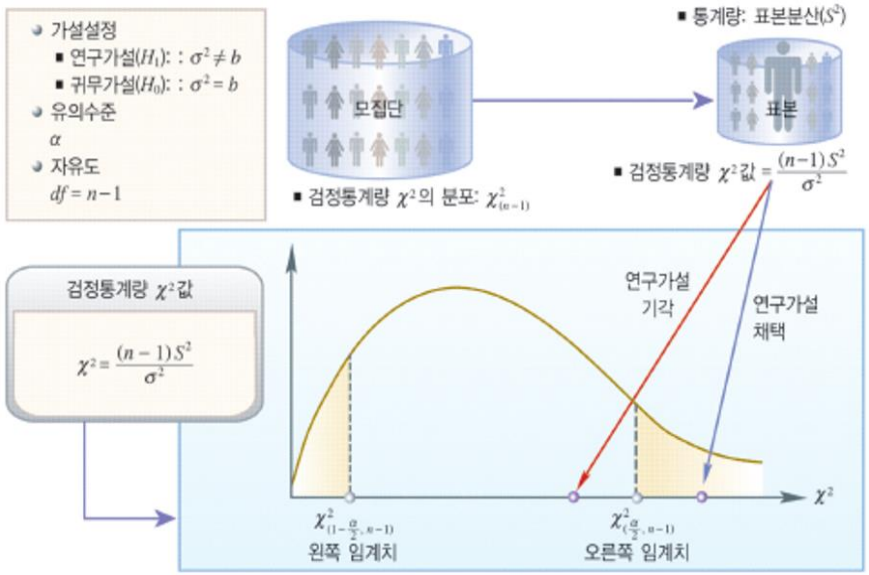
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

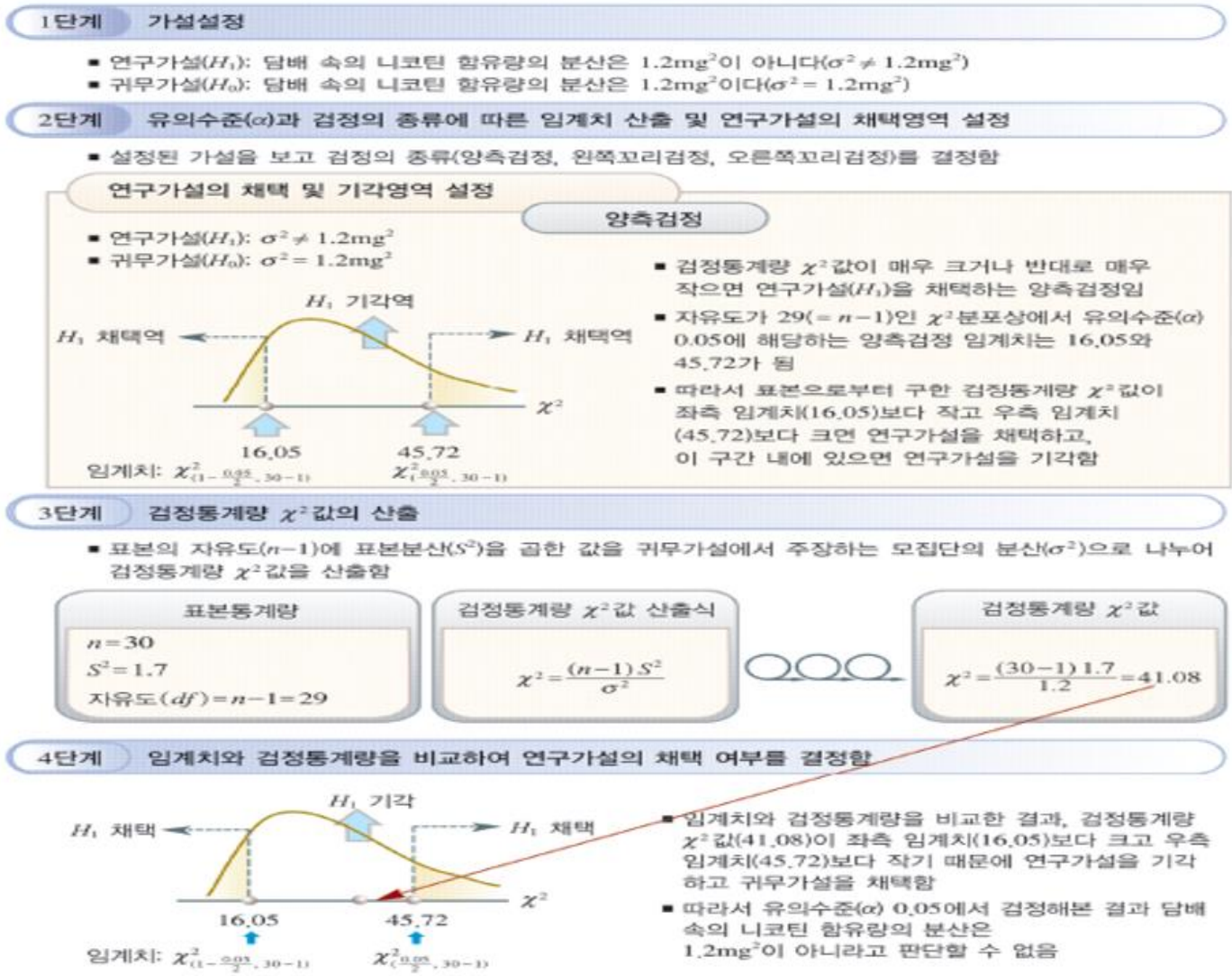
$$F = \frac{s_1^2}{s_2^2}$$



검정(Test): 분산

- χ^2검정





검정(Test): 분산의 차이

- F검정

두 모집단분산에 대한 가설검정

S대학교 e-Marketing 수업은 A반과 B반으로 나누어져 있다. 강좌를 맡으신 교수님은 A반과 B반의 수업 태도를 살펴본 결과 B반보다 A반 학생들의 성적이 더 고를 것이라고 믿고 있다. 과연 그러한지 알아보기 위하여 A반에서 16명, 그리고 B반에서 21명을 무작위로 선정해서 최종 성적을 내어 보니 A반은 평균 78.3에 분산이 37.5이었고, B반은 평균이 76.70이고, 분산은 98.60이었다. 과연 A반의 분포가 B반의 분포보다 평균에 더 밀집되어 성적이 고르다고 볼 수 있는가? 유의수준을 5%로 하여 검정해 보시오.

문제 파악

- 연구가설(H_1): A반 성적의 분산은 B반 성적의 분산보다 작다

  $H_1: \sigma_A^2 < \sigma_B^2$

- 귀무가설(H_0): A반 성적의 분산은 B반 성적의 분산보다 크거나 같다
  혹은 A반 성적의 분산과 B반 성적의 분산은 같다

  $H_0: \sigma_A^2 \geq \sigma_B^2$ 혹은 $\sigma_A^2 = \sigma_B^2$

1단계 가설설정

- 연구가설(H_1): A반 성적의 분산은 B반 성적의 분산보다 작다($\sigma_A^2 < \sigma_B^2$, $\frac{\sigma_A^2}{\sigma_B^2} < 1$, $\frac{\sigma_B^2}{\sigma_A^2} > 1$)
- 귀무가설(H_0): A반 성적의 분산은 B반 성적의 분산보다 크거나 같다($\sigma_A^2 \geq \sigma_B^2$)
  혹은 A반 성적의 분산과 B반 성적의 분산은 같다($\sigma_A^2 = \sigma_B^2$)

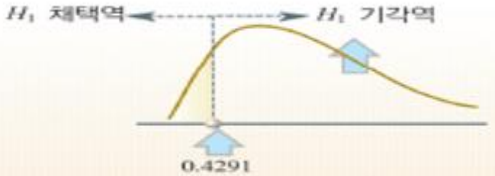
2단계 유의수준(α)과 검정의 종류에 따른 임계치 산출 및 연구가설의 채택영역 설정

- 설정된 가설을 보고 검정의 종류(양측검정, 왼쪽꼬리검정, 오른쪽꼬리검정)를 결정함

연구가설의 채택 및 기각영역 설정

왼쪽꼬리검정

- 연구가설(H_1): $\sigma_A^2 < \sigma_B^2$, $\frac{\sigma_A^2}{\sigma_B^2} < 1$
- 귀무가설(H_0): $\sigma_A^2 \geq \sigma_B^2$ 혹은 $\sigma_A^2 = \sigma_B^2$



- 검정통계량 F값이 임계치보다 작을 경우에 연구가설(H_1)을 채택하는 왼쪽꼬리검정임
- 분자와 분모의 자유도가 각각 15와 20인 F분포상에서 유의수준(α) 0.05에 해당하는 왼쪽꼬리검정의 임계치($F_{(1-0.05;\,20,\,15)}$)를 구하면 0.4291이 됨
- 따라서 표본으로부터 구한 검정통계량 F값이 왼쪽꼬리 임계치(0.4291)보다 작으면 연구가설을 채택하고, 반대로 이보다 크면 연구가설을 기각함

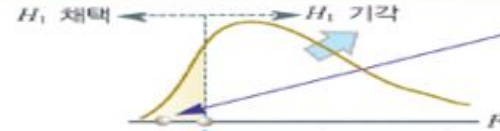
임계치: $F_{(1-\alpha;\,n_A-1,\,n_B-1)} = \frac{1}{F_{(\alpha;\,n_B-1,\,n_A-1)}} = \frac{1}{F_{(0.05;\,20,\,15)}} = \frac{1}{2.33} = 0.4291$

3단계 검정통계량 F값 산출

- 두 표본분산의 비로써, 임계치를 구할 때 사용된 F분포의 분자와 분모의 순서에 맞추어 검정통계량 F값을 산출함

두 모집단의 표본통계량	검정통계량 F값 산출식	검정통계량 F값
$S_A^2 = 37.5$, $n_A = 16$ $S_B^2 = 98.6$, $n_B = 21$ A반의 자유도(df_A): 16−1=15 B반의 자유도(df_B): 21−1=20	$F = \frac{S_A^2}{S_B^2}$	$F = \frac{37.5}{98.6} = 0.38$

4단계 임계치와 검정통계량을 비교하여 연구가설의 채택 여부를 결정함

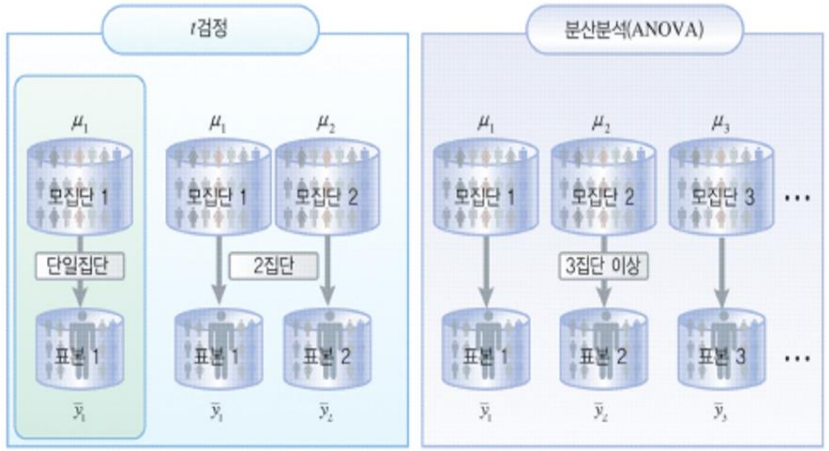


임계치: $F_{(1-0.05;\,15,\,20)} = \frac{1}{F_{(0.05;\,20,\,15)}} = \frac{1}{2.33} = 0.4291$

- 임계치와 검정통계량을 비교한 결과, 검정통계량 F값(0.38)이 왼쪽꼬리 임계치(0.4291)보다 작으므로 귀무가설을 기각하고 연구가설을 채택함
- 따라서 유의수준(α) 0.05에서 검정해본 결과 A반의 분산이 B반의 분산보다 통계적으로 유의하게 작다고 판단할 수 있음

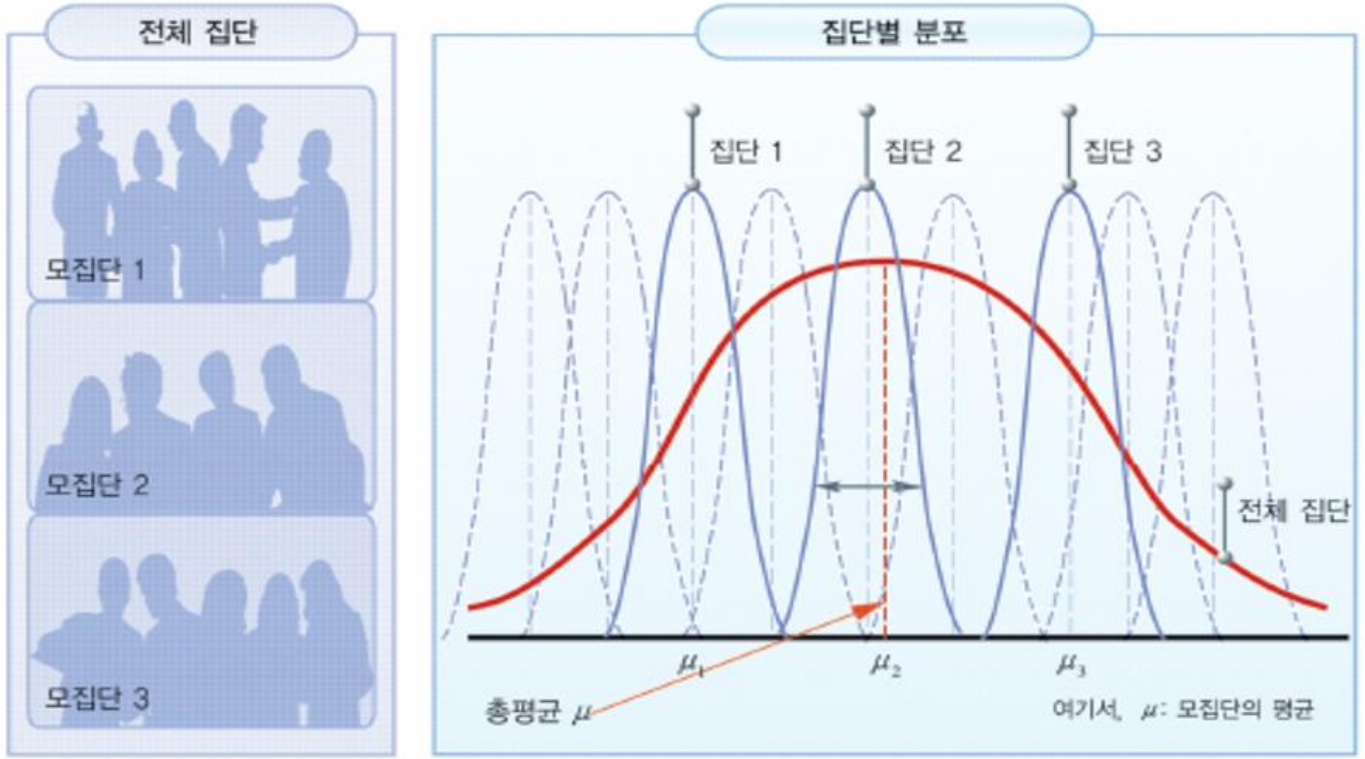
분산분석(ANOVA)

분산분석의 개념 ● 3개 이상의 집단간 평균이 서로 차이가 있는지를 검정하는 분석방법

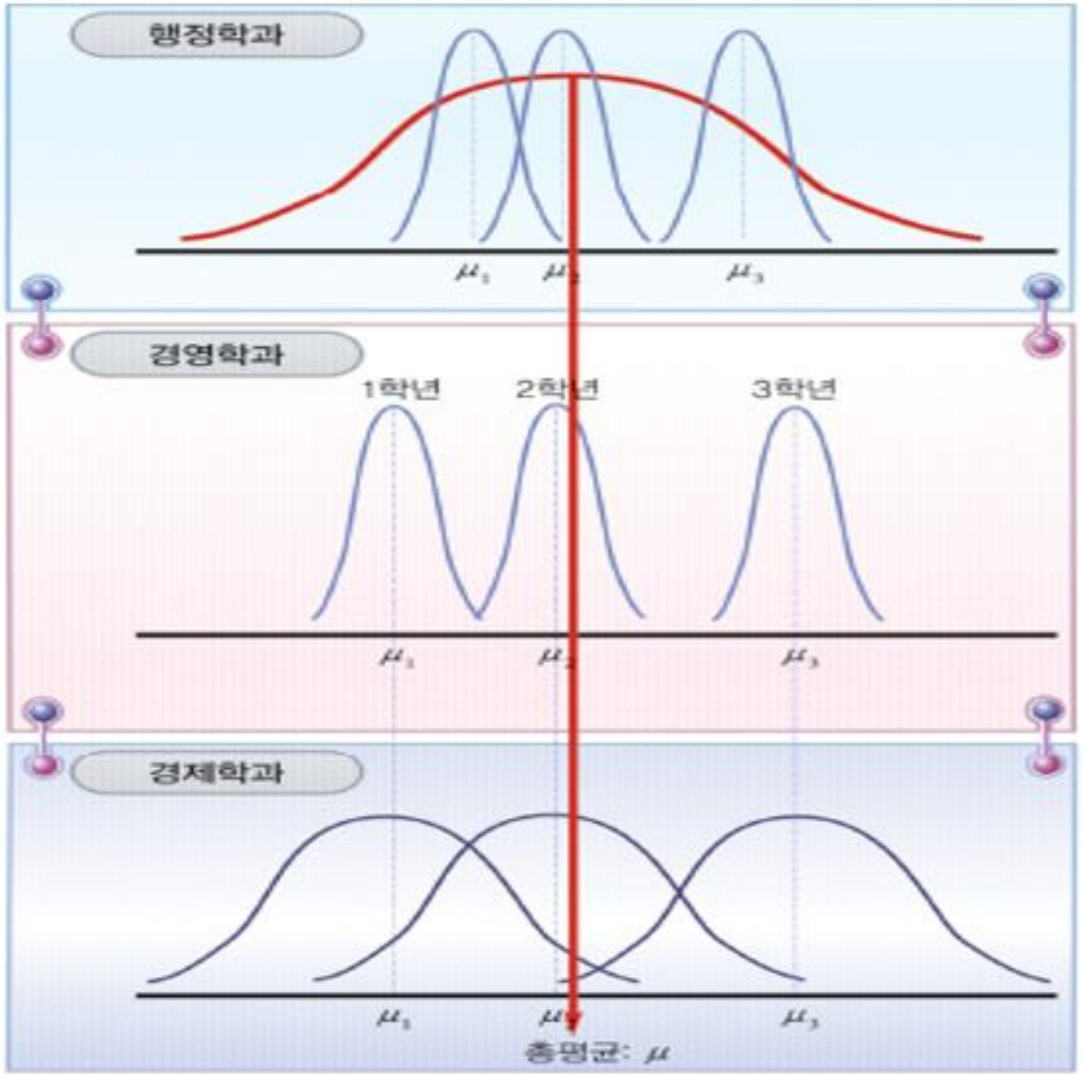


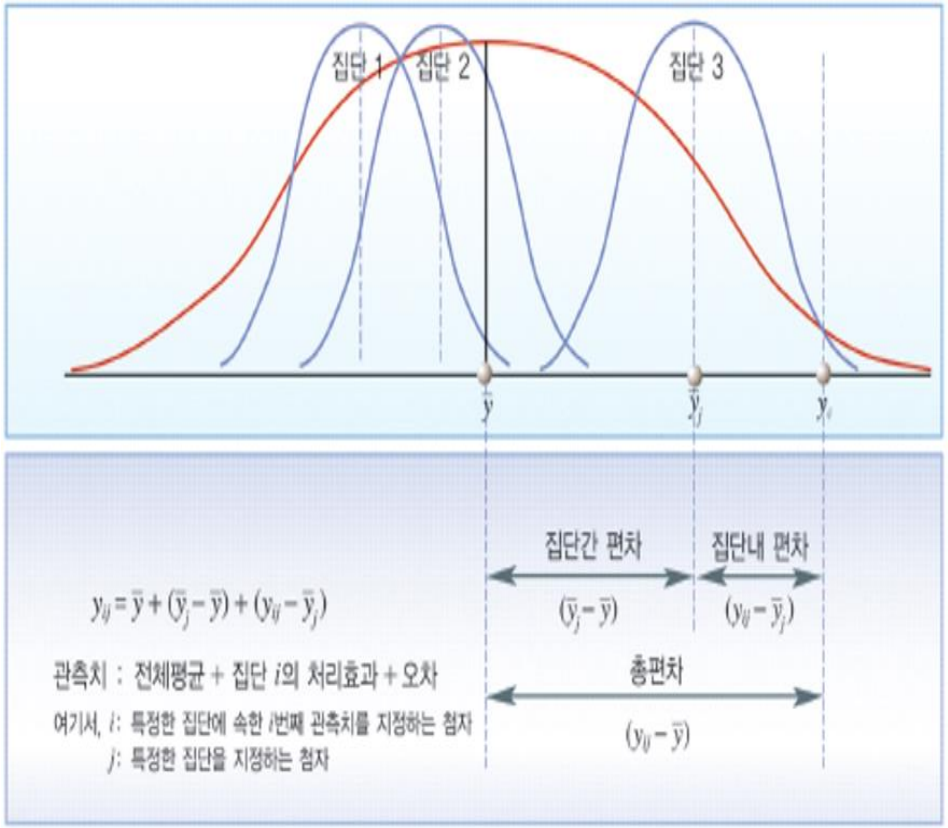
분산분석 ● 집단 간의 평균 차이를 비교하는 데 왜 분산을 이용하여 분석하고, 이를 분산분석이라 하는가?

● 집단의 평균들이 서로 멀리 떨어져 있어서 집단간 평균의 분산이 클수록 집단의 평균들이 서로 다르다고 할 수 있음

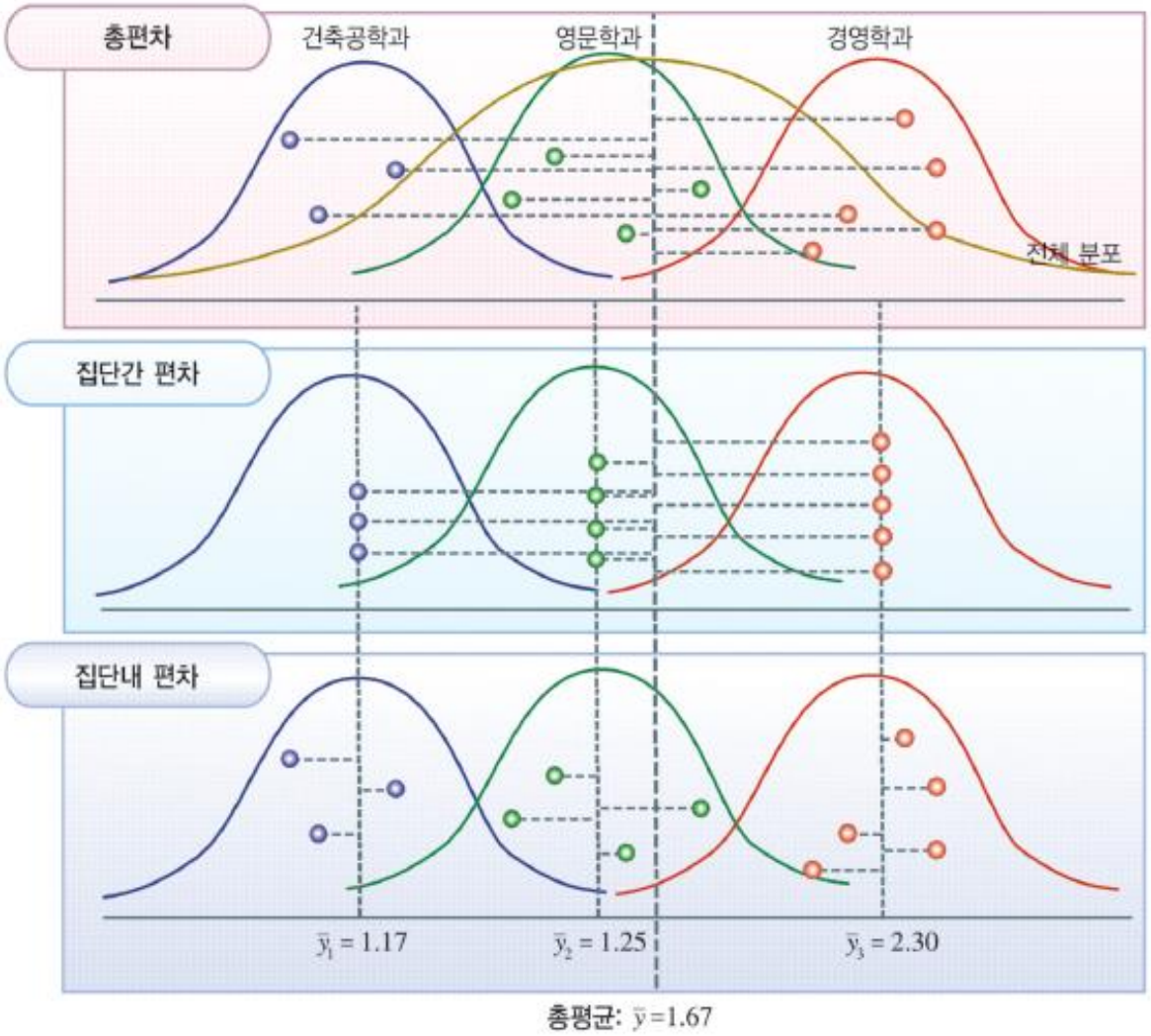


분산분석(ANOVA)


행정학과
경영학과
1학년
2학년
3학년
경제학과
총평균: μ


집단 1
집단 2
집단 3
$y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$
관측치 : 전체평균 + 집단 i의 처리효과 + 오차
여기서, i: 특정한 집단에 속한 i번째 관측치를 지정하는 첨자
j: 특정한 집단을 지정하는 첨자
집단간 편차
$(\bar{y}_j - \bar{y})$
집단내 편차
$(y_{ij} - \bar{y}_j)$
총편차
$(y_{ij} - \bar{y})$

분산분석(ANOVA)



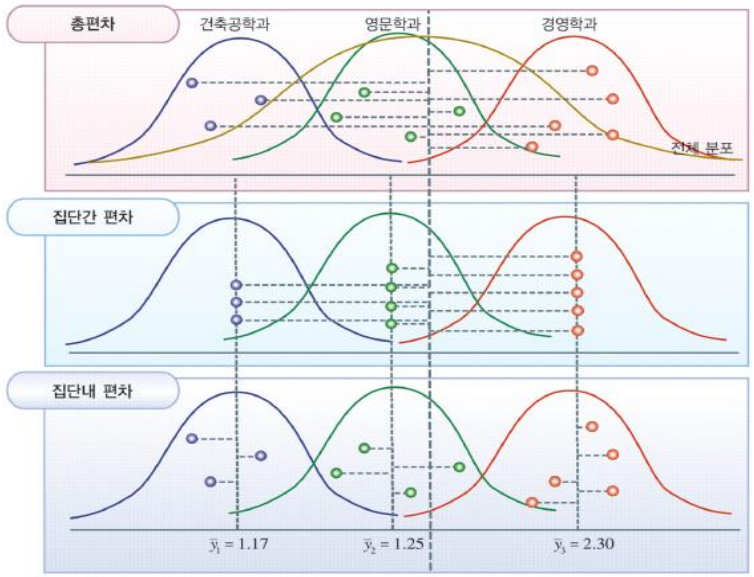
$$\underbrace{(n-1)}_{\text{(총제곱합의 자유도)}} = \underbrace{(g-1)}_{\text{(집단간 제곱합의 자유도)}} + \underbrace{(n-g)}_{\text{(집단내 제곱합의 자유도)}}$$

$$\frac{\sum_{j}^{g}\sum_{i}^{n_j}(y_{ij}-\bar{y})^2}{n-1} \qquad \frac{\sum_{j}^{g}\sum_{i}^{n_j}(\bar{y}_j-\bar{y})^2}{g-1} \qquad \frac{\sum_{j}^{g}\sum_{i}^{n_j}(y_{ij}-\bar{y}_j)^2}{n-g}$$

(총평균제곱: MST) = (집단간 평균제곱: MSB) + (집단내 평균제곱: MSW)

(총분산) (집단간 분산) (집단내 분산)

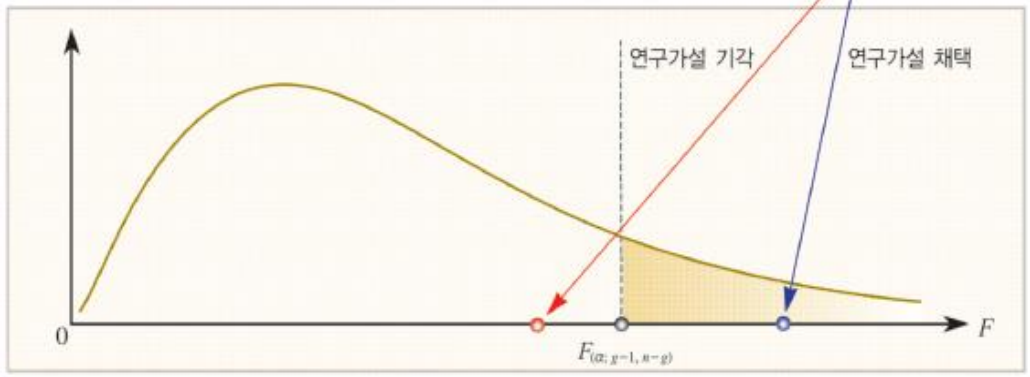
분산분석(ANOVA)



총평균: $\bar{y} = 1.67$

원 천	제곱합(SS)	자유도(df)	평균제곱(MS)	F
집단간	$SSB = \sum_j \sum_i (\bar{y}_j - \bar{y})^2$	$(g-1)$	$MSB = \frac{SSB}{g-1}$	$\frac{MSB}{MSW}$
집단내	$SSW = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$	$(n-g)$	$MSW = \frac{SSW}{n-g}$	
총(합계)	$SST = \sum_j \sum_i (y_{ij} - \bar{y})^2$	$(n-1)$		

집단간 평균제곱(*MSB*) $\frac{\sum_j \sum_i (\bar{y}_j - \bar{y})^2}{g-1}$ ÷ **집단내 평균제곱(*MSW*)** $\frac{\sum_j \sum_i (y_{ij} - \bar{y}_j)^2}{n-g}$ = **검정통계량 *F*값** $F = \frac{MSB}{MSW}$



- 분산분석은 집단간 평균제곱(*MSB*)을 집단내 평균제곱(*MSW*)으로 나눈 통계량 *F*값을 검정통계량값으로 하여 집단간 평균의 차이가 통계적으로 유의한지를 분석함
- 연구가설(H_1): 비교하려는 집단들의 평균이 모두 같지는 않음
  - 적어도 한 집단의 평균은 나머지와 차이가 있음

분산분석(ANOVA)

주효과 검정만이 가능한 이원분산분석

- 2개의 독립변수(명목변수)에 의하여 구분되는 각 집단 내에 오직 1개씩의 관측치만 있는 경우 – 주효과(main effect) 분석만이 가능함

- 매장의 규모(요인 i)와 지리적 위치(요인 j)에 따른 매출액 차이 분석

y_{ij} $\bar{y}_{i.}$

규모 \ 위치	강남	강북	요인 i(평균)
대형	10	7	8.50
중형	8	3	5.50
소형	4	1	2.50
요인 j(평균)	7.33	3.67	5.50

$\bar{y}_{j}$ $\bar{y}_{..}$

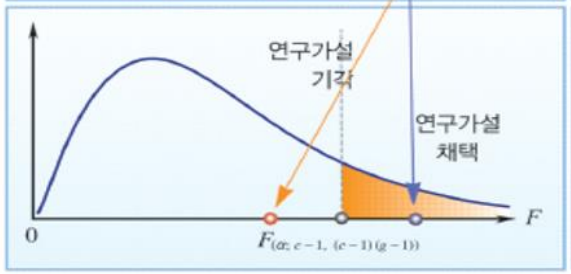
여기서, i: 요인 1을 요인 i로 표기함
j: 요인 2를 요인 j로 표기함
y_{ij}: 요인 i와 요인 j에 의해 구분되는 집단에 속한 관측치
$\bar{y}_{..}$: 전체평균
$\bar{y}_{i.}$: 요인 i에 의해서만 구분되는 집단의 평균
$\bar{y}_{j}$: 요인 j에 의해서만 구분되는 집단의 평균

- 매장규모(요인 i)와 지리적 위치(요인 j)로 구분되는 세분집단(셀) 내에 오직 1개씩의 관측치만 있으므로 상호작용효과 분석이 불가능함

원 천	제곱합(SS)	자유도(df)	평균제곱(MS)	F 비
요인 i	SSB_i	$c-1$	MSB_i	$\frac{MSB_i}{MSE}$
요인 j	SSB_j	$g-1$	MSB_j	$\frac{MSB_j}{MSE}$
오차	SSE	$(c-1)(g-1)$	MSE	
총(합계)	SST	$cg-1$		

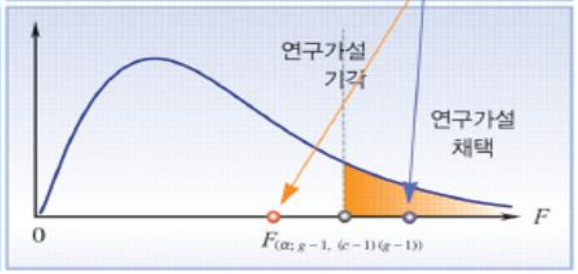
- F분포: $F_{c-1,\,(c-1)(g-1)}$
- 유의수준: α
- 임계치: $F_{\alpha;\,c-1,\,(c-1)(g-1)}$

요인 i의 검정통계량 F값 $F = \frac{MSB_i}{MSE}$



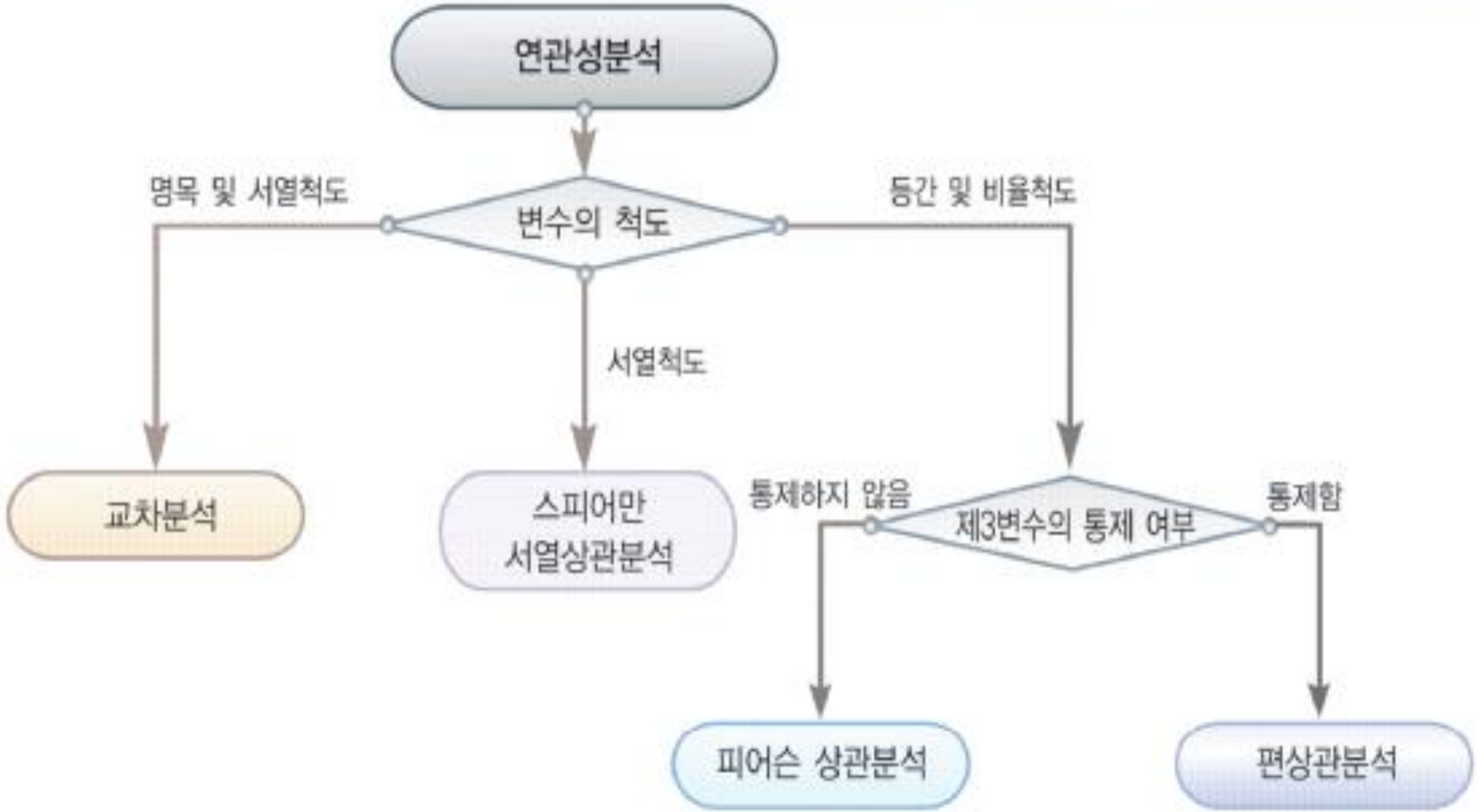
- F분포: $F_{g-1,\,(c-1)(g-1)}$
- 유의수준: α
- 임계치: $F_{\alpha;\,g-1,\,(c-1)(g-1)}$

요인 j의 검정통계량 F값 $F = \frac{MSB_j}{MSE}$



연관성 분석

연관성분석 ● 2개 변수들 간의 연관성을 파악하는 분석방법



연관성 분석

구분	척도	분석방법	사용 통계량	기타 변수 개입여부
상관분석	등간척도 비율척도	편상관분석	$r_{xy\cdot z} = \frac{r_{xy} - (r_{xz})(r_{yx})}{\sqrt{1 - {r_{xz}}^2}\sqrt{1 - {r_{yz}}^2}}$	○
		(피어슨) 상관분석	$r = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$	×
	서열척도	스피어만 상관분석	$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2)}{n(n^2-1)}$	-
교차분석	명목척도	교차분석	χ^2검정	-

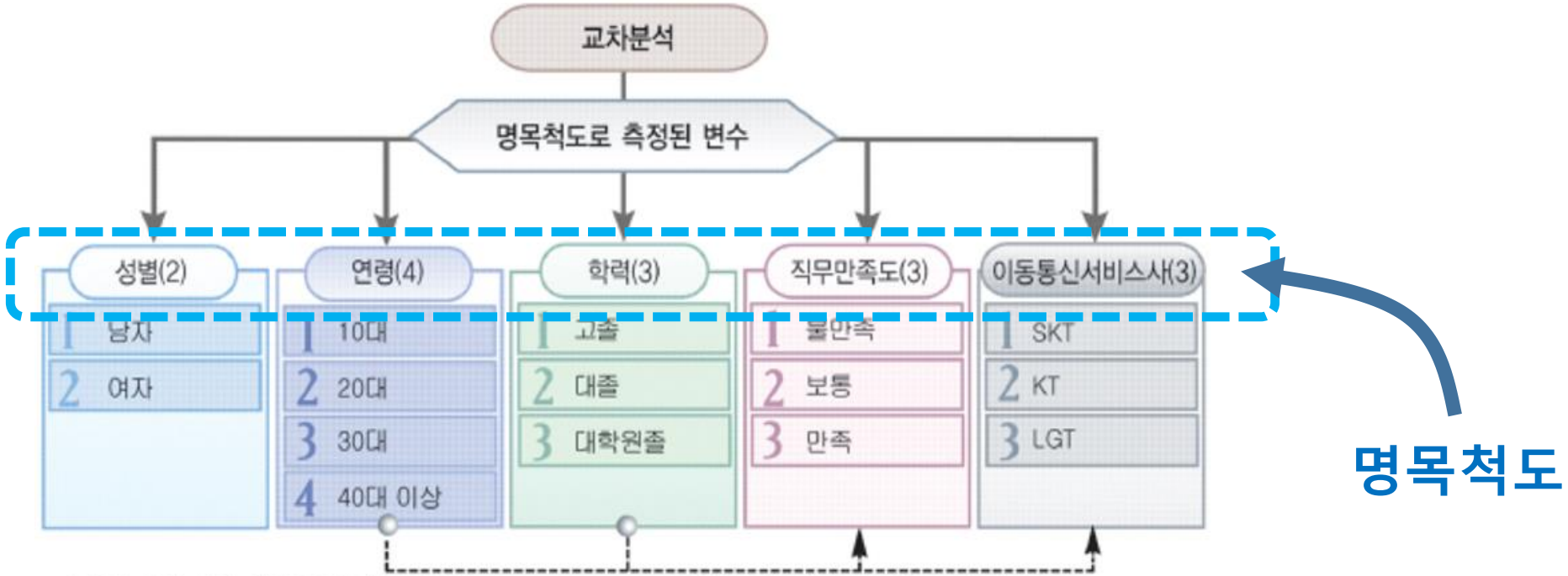
적합도 검정

예) 주사위를 120회 던지는 실험을 실시하여 나온 결과이다. 이 주사위는 정상적인가?

	1	2	3	4	5	6
관측치	18	24	17	23	16	22
기대치	20	20	20	20	20	20

교차분석

- 명목(서열)척도로 측정된 **두 변수 간**의 상호연관성을 알아보기 위한 분석



공분산(Covariance)

ID	국어	영어
1	95	95
2	90	95
3	80	75
4	60	70
5	40	35
6	80	80
7	95	90
8	30	25
9	15	10
10	60	70
평균	64.5	64.5
분산	808	925
공분산	762	

ID	국어	영어
1	9.5	9.5
2	9.0	9.5
3	8.0	7.5
4	6.0	7.0
5	4.0	3.5
6	8.0	8.0
7	9.5	9.0
8	3.0	2.5
9	1.5	1.0
10	6.0	7.0
평균	6.45	6.45
분산	8.08	9.25
공분산	7.62	

✓ $Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

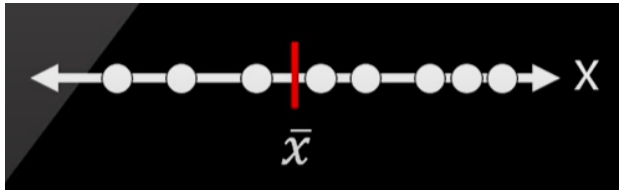
✓ 공분산: X의 편차와 Y의 편차를 곱한 값들의 평균으로, **데이터**에 따라 공분산 **크기가 달라진다**.

- Cov(X,Y) > 0 : X↑ , Y↑
- Cov(X,Y) < 0 : X↑ , Y↓
- Cov(X,Y) = 0 : No linear relationship

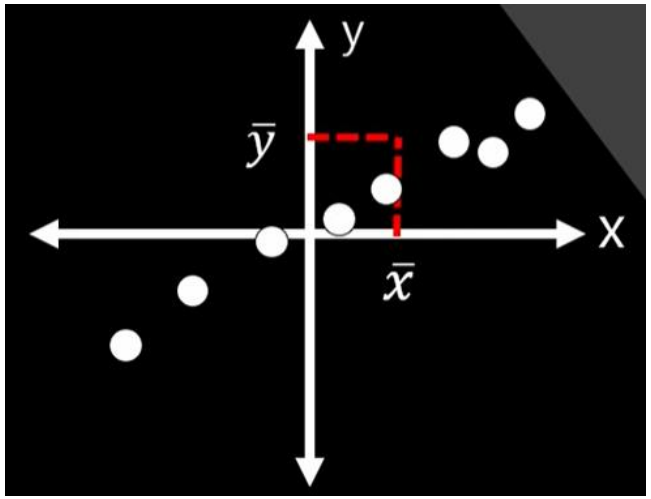
✓ 공분산은 두 변수 간에 **양의 상관**관계가 있는지? **음의 상관**관계가 있는지? 정도만 알려줌

✓ 상관관계가 얼마나 큰지는 제대로 반영하지 못함

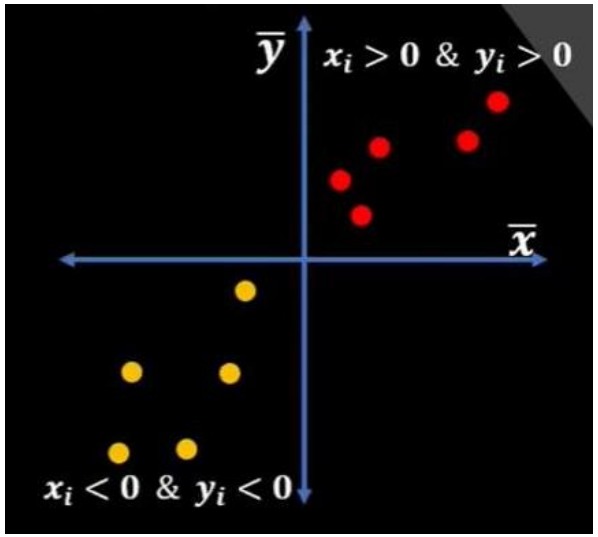
분산 & 공분산



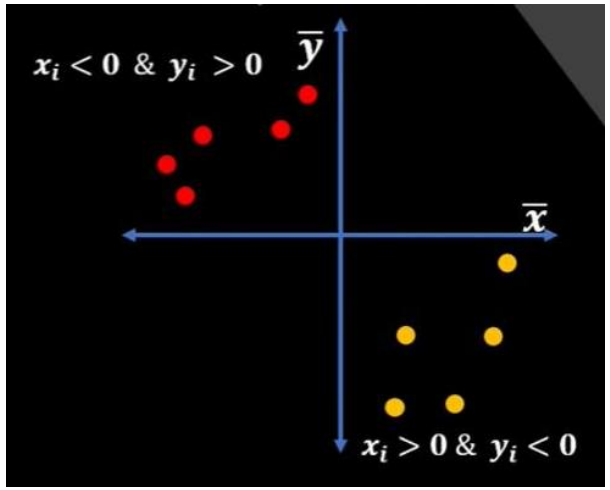
$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$



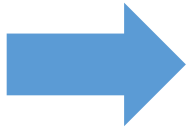
$Cov(x,y) = +$



$Cov(x,y) = -$

상관계수(Correlation Coefficient)

ID	국어	영어
1	95	95
2	90	95
3	80	75
4	60	70
5	40	35
6	80	80
7	95	90
8	30	25
9	15	10
10	60	70
평균	64.5	64.5
분산	808	925
공분산	762	



ID	국어	영어
1	9.5	9.5
2	9.0	9.5
3	8.0	7.5
4	6.0	7.0
5	4.0	3.5
6	8.0	8.0
7	9.5	9.0
8	3.0	2.5
9	1.5	1.0
10	6.0	7.0
평균	6.45	6.45
분산	8.08	9.25
공분산	7.62	

$$\gamma = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{762}{\sqrt{808}\sqrt{925}} = 0.88$$

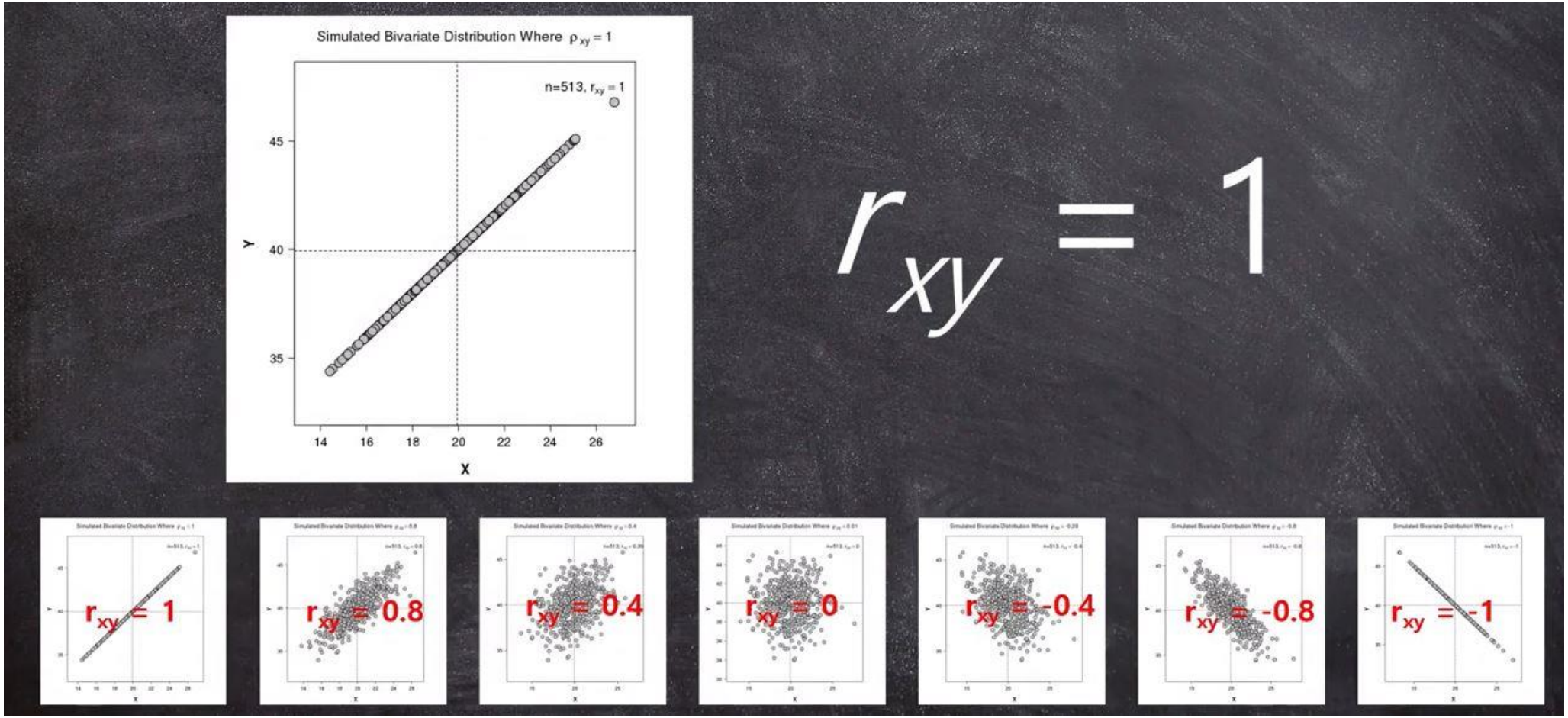
$$\gamma = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{7.62}{\sqrt{8.08}\sqrt{9.25}} = 0.88$$

상관계수(Correlation Coefficient)

- $$\gamma = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$$

- 상관계수: **공분산을 표준화한 값**
- 절대값은 1보다 작거나 같음
- X, Y가 완벽한 선형관계이면, γ =1 or -1
- 상관계수가 1 또는 -1에 근접할수록 힘(power)가 크다는 것
- 여기서의 힘이란 점들이 모여있는 정도

상관계수(Correlation Coefficient)



내적 & 상관계수

$$Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = E[XY] - \mu_X\mu_Y$$

$$= \frac{1}{N}\sum_{i=1}^{N} X_iY_i - \mu_X\mu_Y$$

$$= \frac{1}{N} < X,Y > - \mu_X\mu_Y$$

$$\boldsymbol{\rho} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)$$

$$if\ \mu_X\ \&\ \mu_Y = 0$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{X_iY_i}{\sigma_X\sigma_Y} = \frac{< X,Y >}{\sqrt{\sum X^2}\sqrt{\sum Y^2}} = \frac{< \boldsymbol{X},\boldsymbol{Y} >}{\|\boldsymbol{X}\|\|\boldsymbol{Y}\|}$$

$$= cos\theta$$

- ✓ 내적 $a \cdot b$ (a^Tb, $<a,b>$) 결과 행렬로부터 공분산을 구하기 위해서는 원래의 data matrix에서 평균이 zero가 되도록 **표준화** 해주어야 한다.
- ✓ 상관계수 $\boldsymbol{\rho}$ 는 두 벡터 간의 $cos\theta$ 값이 되기 때문에 $-1 \le \boldsymbol{\rho} \le 1$ 갖게 된다.
- ✓ **표준화** 한 경우에는, 상관계수 $\boldsymbol{\rho}$ = 내적의 $\boldsymbol{cos\theta}$ 성립하며,

$$\boldsymbol{Cov(X,Y) = < X,Y >}$$