강원대학교
AI 소프트웨어학과

# 머신러닝1

## – 기초통계 –

### 추론통계(분산의 분포– ANOVA)

KNU
강원대학교

## F-test 란?

- F-test는 두 집단의 분산이 동일한지를 비교하는 통계적 검정
- 초기에는 각 집단의 평균의 분산(Between-group Variance) → 집단 간의 차이
  - 각 집단의 평균값이 멀리 떨어져 있음 → A집단은 B집단과 차이가 존재함
  - A의 평균으로 부터 B의 평균 까지의 분산 → Between-group Variance → 집단 간의 차이

    **→ 따라서 분산분석이라 부름(두 집단의 분산이 동일한지 검정)**

    **→ 등분산 가정 여부를 확인할 때 사용(분산의 동질성 검정)**

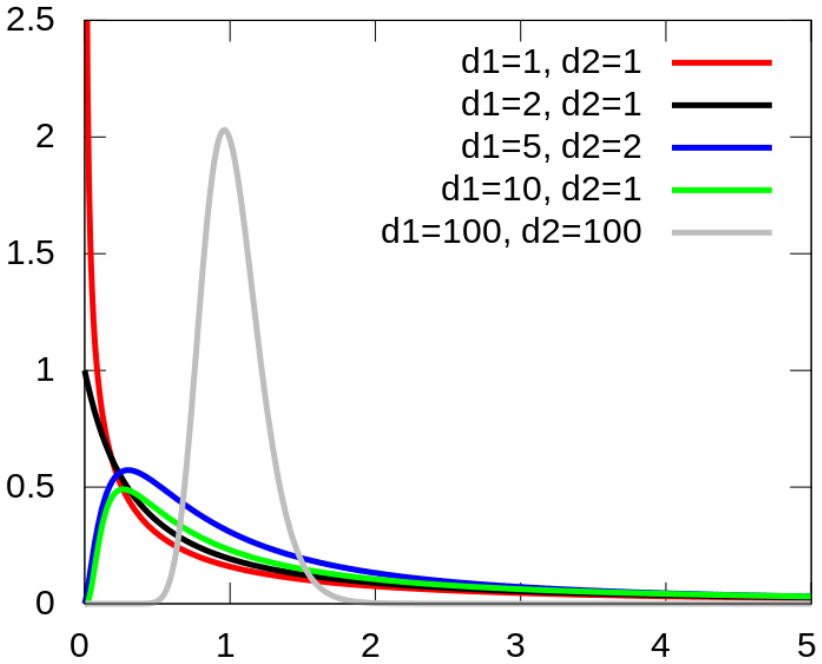
$$F = \frac{s_1^2}{s_2^2}$$

| 학원 | 점수 | 평균 | 분산 |
|---|---|---|---|
| A | 75, 76, 74, 75, 76 | 75.2 | 0.7 |
| B | 80, 81, 79, 80, 81 | 80.2 | 0.7 |
| C | 85, 86, 84, 85, 86 | 85.2 | 0.7 |
| D | 66,95,64,95,65 | 77 | 270.5 |

## 추론통계

### F-분포란?

- F-분포(Fisher-Snedecor distribution) : F-value는 집단의 분산의 비율
- X축은 두 분산의 비율, Y축은 F값이 발생할 확률 밀도 함수 → F-분포도 양수를 가짐
- 집단의 데이터가 동일한 분포를 따르고, 각 집단의 데이터가 서로 독립일 경우 → 자유도가 커질수록 F-분포는 대칭성이 증가함
- 비교하는 집단들 간의 표본은 서로 독립이어야 하며, 이는 한 그룹의 측정값이 다른 그룹의 측정값에 영향을 주지 않아야 함을 의미




$$F = \frac{\chi_1^2/d1}{\chi_2^2/d2} = \frac{s_1^2}{s_2^2}$$

두 카이제곱 분포(=분산 정보)를 서로 비교

서로 크기가 다를 수 있으니 자유도로 나눠 평균 분산을 만든 후 비율 계산

## F-test 란?

- F-test는 두 집단의 분산이 동일한지를 비교하는 통계적 검정
- 각 집단의 평균 까지의 분산(Between-group Variance) → 집단 간의 차이
  - 각 집단의 평균값이 서로 떨어져 있음 → A집단은 B집단과 차이가 존재함
  - 기준으로 부터 각각의 평균의 분산 → Between-group Variance → 집단 간의 차이

## 분산분석(F-test)

- 가설 설정
- $H_0$ : 두집단의 분산은 같다. → 모집단의 분산 = 표본집단의 분산
- $H_1$ : 두집단의 분산은 차이가 있다. → 모집단의 분산 $\neq$ 표본집단의 분산

## 추론통계

### 분산분석(F-test)

- 가설 설정
- $H_0$ : 두집단의 분산은 같다.
- $H_1$ : 두집단의 분산은 차이가 있다.

**예제)**

Set A: {10.1, 10.2, 10.3, 10.0, 10.1, 10.2, 10.3, 10.0, 10.1, 10.2}
Set B: {9.8, 10.5, 10.2, 9.7, 10.4, 10.3, 9.6, 10.6, 9.9, 10.7}

## 분산분석(F-test)

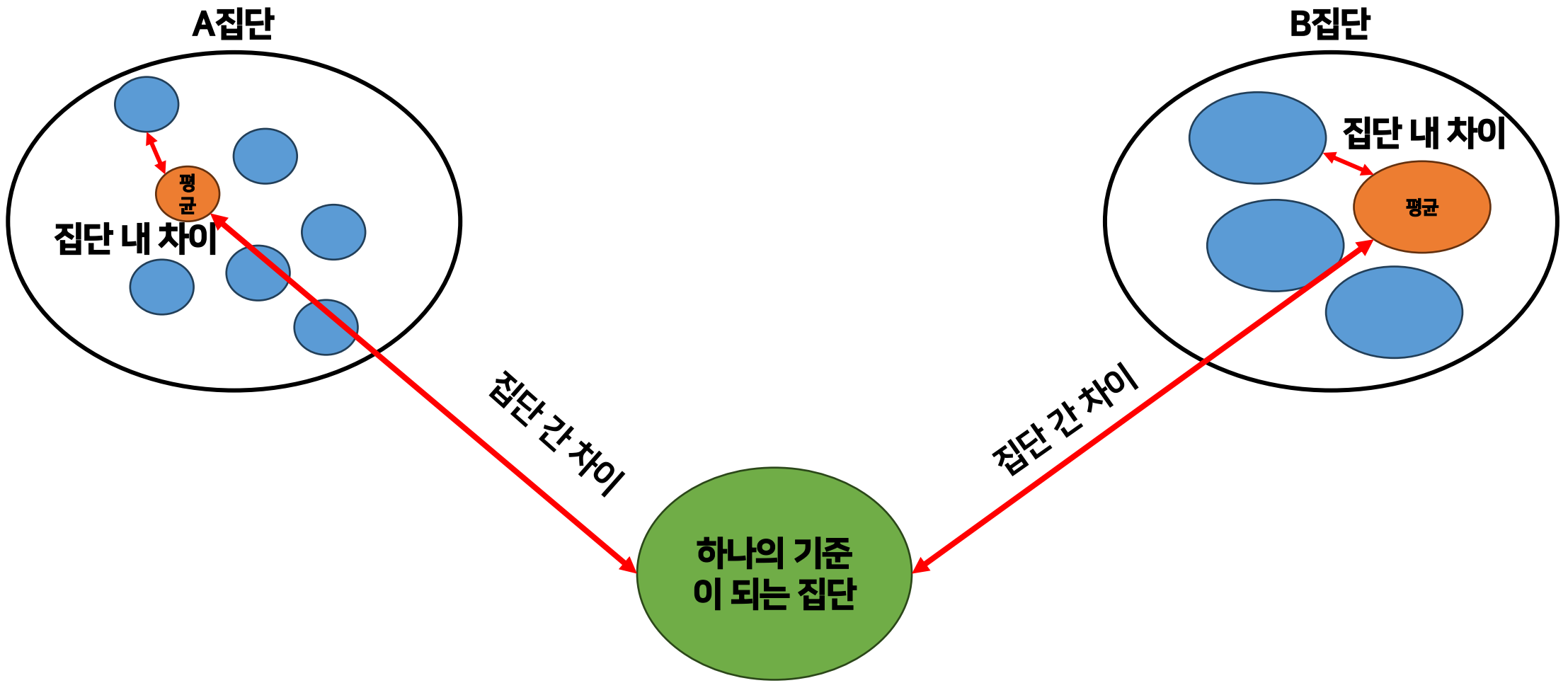
```
set_a <- c(10.1, 10.2, 10.3, 10.0, 10.1, 10.2, 10.3, 10.0, 10.1, 10.2)  # 분산 작음
set_b <- c(9.8, 10.5, 10.2, 9.7, 10.4, 10.3, 9.6, 10.6, 9.9, 10.7)     # 분산 큼

# F-test: 분산 동질성 검정 (양측)
result <- var.test(set_a, set_b)
print(result)

# F-test: 분산 동질성 검정 (단측) : 귀무가설 기각 시 a집단은 b집단보다 분산이 작다.
var.test(set_a, set_b, alternative = "less")

# F-test: 분산 동질성 검정 (단측) : 귀무가설 기각 시 a집단은 b집단보다 분산이 크다.
var.test(set_a, set_b, alternative = "greater")
```

## 분산분석(F-test)

```
df=read.csv("C:/Users/tiock/Desktop/Machine_set.csv")

set_a <- subset(df, machine == "A")$value
set_b <- subset(df, machine == "B")$value

# F-test: 분산 동질성 검정 (양측)
result <- var.test(set_a, set_b)
print(result)

# F-test: 분산 동질성 검정 (단측) : 귀무가설 기각 시 a집단은 b집단보다 분산이 작다.
var.test(set_a, set_b, alternative = "less")

# F-test: 분산 동질성 검정 (단측) : 귀무가설 기각 시 a집단은 b집단보다 분산이 크다.
var.test(set_a, set_b, alternative = "greater")
```

## 추론통계

### 분산분석(F-test)



## 추론통계

### 분산분석(F-test)

```
set_a <- c(10.2, 10.3, 10.1, 10.0, 10.3, 10.2, 10.1, 10.2, 10.1, 10.3) #기준이 되는 분산 값
set_b <- c(10.1, 10.0, 10.2, 10.3, 10.0, 10.1, 10.2, 10.1, 10.2, 10.1)
set_c <- c(10.2, 10.3, 10.3, 10.3, 10.4, 10.3, 10.2, 10.4, 10.3, 10.4)


var.test(비교대상, 기준, alternative = "less")
```

- 가설 설정
- $H_0$ : 모집단의 분산 = 표본집단의 분산
- $H_1$ : 모집단의 분산 ≠ 표본집단의 분산

$$F = \frac{\text{비교대상}}{\text{기준}} = \frac{s_1^2}{s_2^2}$$

**예제)**

기준 부품의 규격이 A에서 B,C 부품의 생산 규격은 차이가 있는가?

## 추론통계

### 분산분석(F-test)

- 가설 설정
- $H_0$ : 모집단의 분산 = 표본집단의 분산
- $H_1$ : 모집단의 분산 ≠ 표본집단의 분산

**예제)**

부품의 규격이 A={10.2,10.3,10.1,10.0,10.3,10.2,10.1,10.2,10.1,10.3}일 때, 이 부품의 분산 보다 크거나 작으면 잘못된 설계

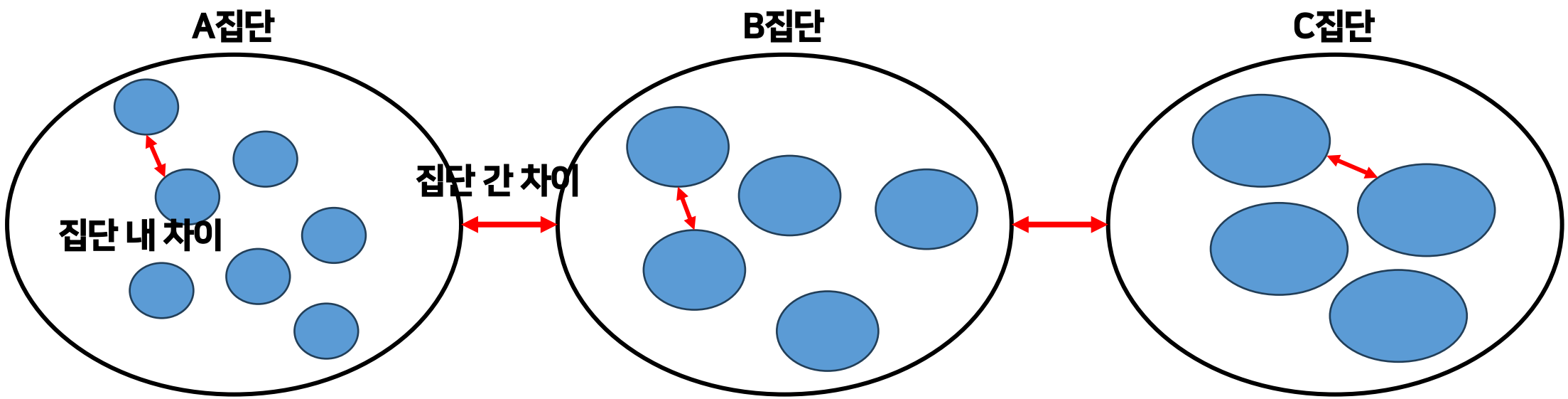
- $H_0$ : 모집단의 분산 = Set ?의 분산($H_0$ : $\sigma^2$ = Set B or C)
- $H_1$ : 모집단의 분산 ≠ Set ?의 분산($H_1$ : $\sigma^2$ < Set B or C, $\sigma^2$ > Set B or C)

**예제)**

Set A: {10.2, 10.3, 10.1, 10.0, 10.3, 10.2, 10.1, 10.2, 10.1, 10.3}
Set B: {10.1, 10.0, 10.2, 10.3, 10.0, 10.1, 10.2, 10.1, 10.2, 10.1}
Set C: {10.2, 10.3, 10.3, 10.3, 10.4, 10.3, 10.2, 10.4, 10.3, 10.4}

## 추론통계

### 분산분석(F-test)



## 추론통계

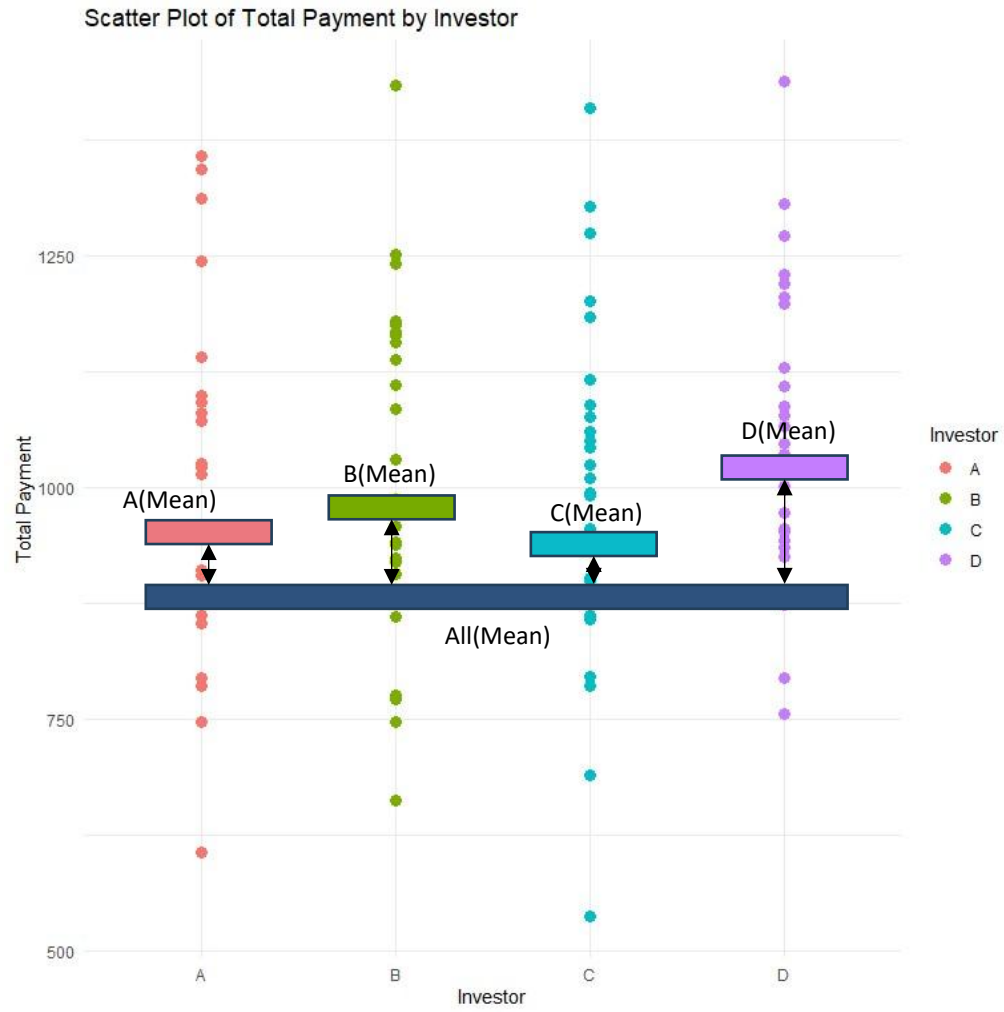
### 가설 검정 및 추론통계

- 두 연속 집단의 평균 차이 비교(T-Test, Z-Test)
- 두 명목 집단의 연관성 비교(chi square-Test)
- 두 연속 집단의 분산 차이 비교(F-Test)

- 그룹을 비교할 때 세개 이상의 그룹을 비교할 수는 없을까?
- F-검정을 사용하는 분산분석(ANOVA)은 세 개 이상의 그룹의 평균 차이를 검정하는 방법
- ANOVA는 F-검정을 통해 그룹 간 평균 차이를 검정하는 방법으로, 집단 간 변동(분산)이 집단 내 변동(분산)보다 충분히 크면 평균에 유의한 차이가 있다고 판단하는 방법

## 추론통계

### 분산분석(F-test)

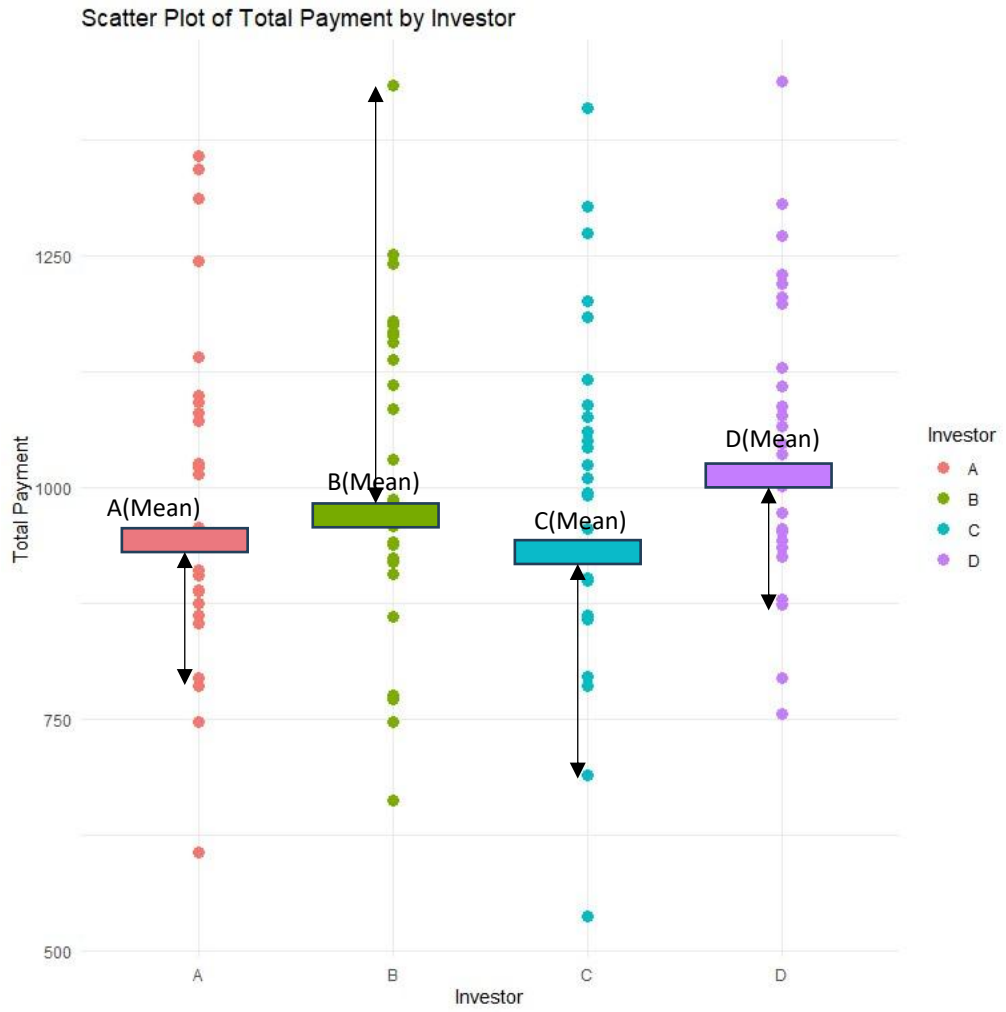



- F-값은 두 가지의 분산의 비율이므로 두 가지의 평균이 필요함
- 변수는 하나이고, 이 안에 속하는 집단이 4개
- 두 가지의 분산을 구할 수 있음
- Between Variance : 전체평균에서 각 집단의 평균이 얼마나 멀리 떨어져 있나?

| 수치값 | 집단 |
|---|---|
| 10 | A |
| 20 | B |
| 25 | C |
| 21 | D |
| 26 | B |

## 추론통계

### 분산분석(F-test)




- F-값은 두 가지의 분산의 비율이므로 두 가지의 평균이 필요함
- 변수는 하나이고, 이 안에 속하는 집단이 4개
- 두 가지의 분산을 구할 수 있음
- Between Variance : 전체평균에서 각 집단의 평균이 얼마나 멀리 떨어져 있나?
- Within Variance : 각 집단의 데이터가 각 집단의 평균에서 얼마나 떨어져 있나?
- Between Variance가 Within Variance보다 크면 → 적어도 어느 한 그룹의 평균값이 전체 평균과는 다르다고 할 수 있음

## 추론통계

### 분산분석(ANOVA)

- 집단 간 분산 (Between-group variance) → 각 집단의 평균이 전체 평균으로부터 얼마나 떨어져 있는지를 나타냄
  → 전체 값의 평균을 기준으로 집단 간 평균 차이를 측정

- 집단 내 분산 (Within-group variance) → 각 집단 내부의 개별 값들이 자기 집단의 평균으로부터 얼마나 떨어져 있는지를 나타냄
  → 각집단의 평균을 기준으로 집단 내부의 값들의 차이를 측정

- 집단 간 분산이 집단 내 분산보다 크면 → 그룹 간 평균 차이가 우연이 아닌 실제 차이일 가능성이 높음
  → 집단 간의 차이가 집단 내의 차이보다 커야 두 집단이 실제로 다른 집단이라고 할 수 있음
  → 따라서 귀무가설($H_0$ : 모든 집단 평균에는 차이가 없다.)을 기각
  → 대립가설($H_1$ : 적어도 하나의 집단의 평균은 차이가 있다.)을 채택

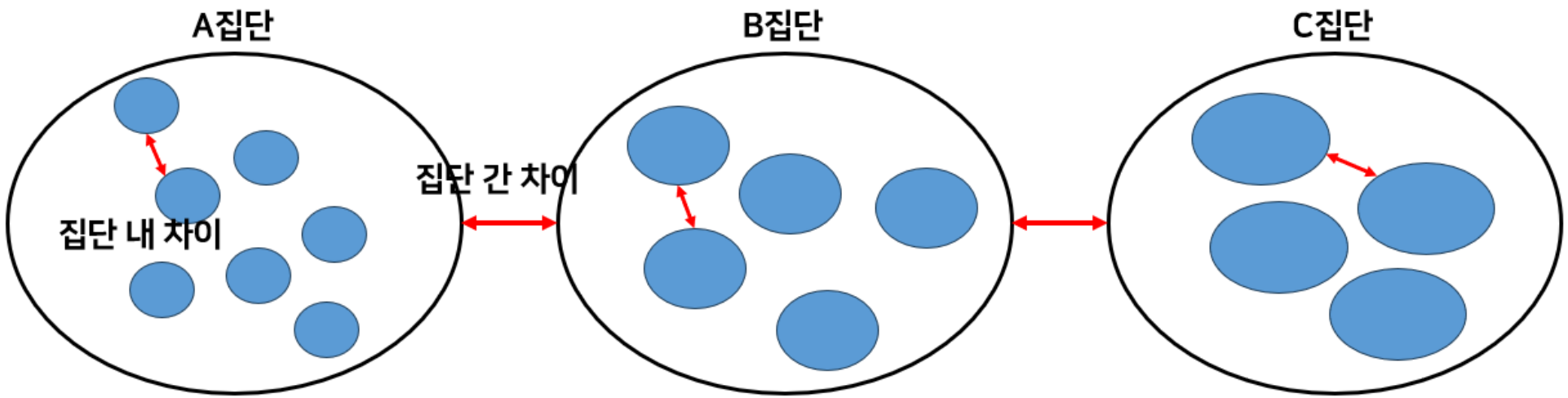
- 집단 내 분산이 집단 간 분산보다 크면 → 관측된 그룹 간 평균 차이는 우연일 가능성이 높음(집단 간 차이를 밝히는 데는 유의한 결과를 얻지 못함)

## 추론통계

### 분산분석(ANOVA)

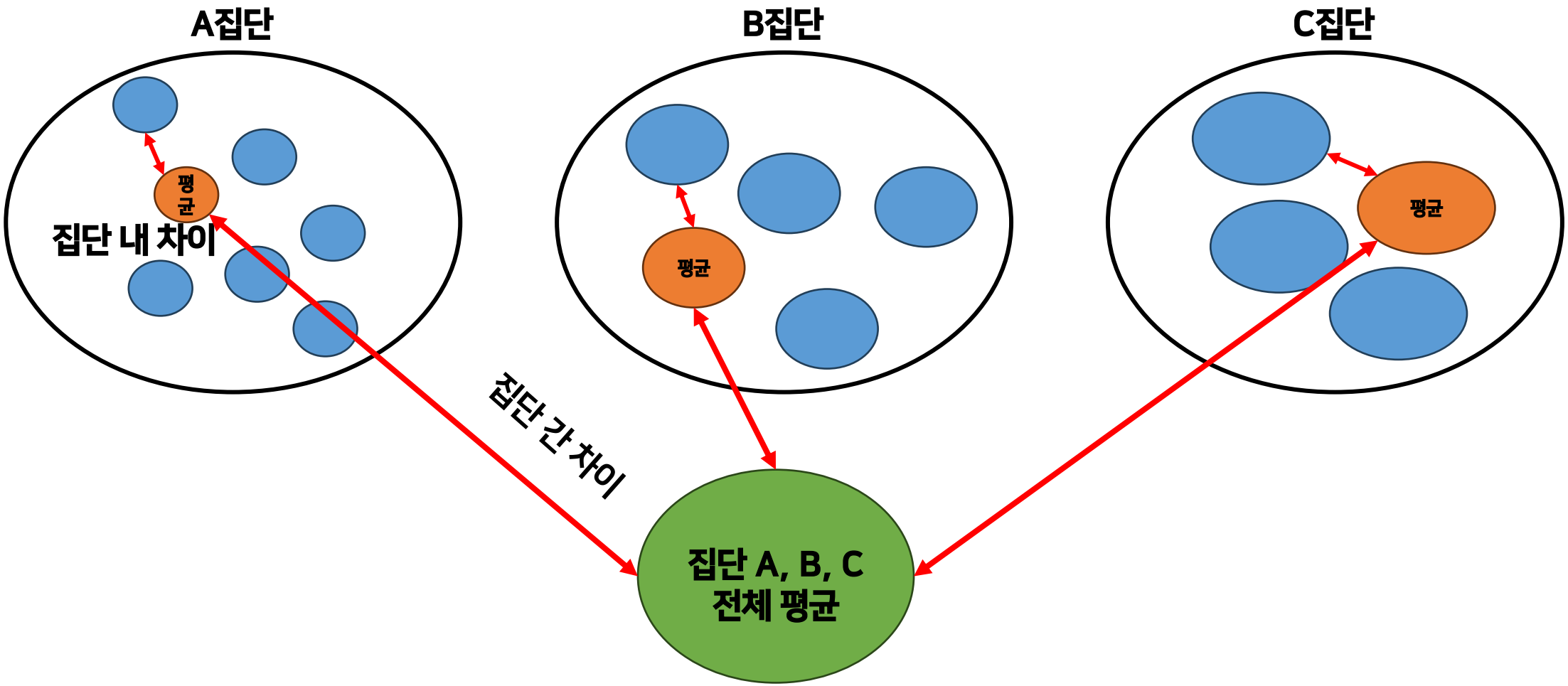
- 분산분석(ANOVA) : 세 개 이상의 그룹 평균을 비교하여 평균의 차이가 존재하는지 판단하는 방법
- 집단 간 분산을 집단 내 분산으로 나눈 것 → 분산을 활용해 평균을 비교
  - F = 1 → 집단 간 분산 = 집단 내 분산 → 평균 차이 없음
  - F > 1 → 집단 간 분산 > 집단 내 분산 → 평균 차이 있음 (의심)
  - F < 1 → 집단 간 분산 < 집단 내 분산 → 집단 평균 간 차이가 없고, 오히려 집단 내 불확실성이 더 큼

$$F = \frac{\text{집단 간 평균 제곱}(MSB)}{\text{집단 내 평균제곱}(MSW)}$$



## 추론통계

### 분산분석(ANOVA)



## 추론통계

### 분산분석(ANOVA)

- 집단 간 분산(Between-group Variance)을 집단 내 분산(Within-group Variance)으로 나눈 것 → 분산을 활용해 평균을 비교
- MSB : Mean Square Between → 집단 간의 평균 차이를 반영, Sum of Squares Between Groups($SS_{between}$)
- MSW : Mean Square Within → 각 그룹 내부의 변동성, Sum of Squares Within Group($SS_{within}$)

$$F = \frac{Between\ Variance}{Within\ Variance} = F = \frac{MS_{Between\ Group}}{MS_{within\ Group}} = F = \frac{SS_{between}/(k-1)}{SS_{within}/(N-k)}$$

$k =$ 집단의 수, $N =$ 전체 데이터수

$$SS_{between} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{total})^2 = \sum(\text{각 그룹의 데이터 개수} \times (\text{각 그룹 평균} - \text{전체 평균}))^2$$

$$SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 = \sum(\text{그룹 내의 각 데이터 값} - \text{해당 그룹의 평균})^2$$

$n_i =$ 집단 $i$의 데이터수
$j$ : 각 그룹안의 개별 관측값 번호

## 추론통계

### 분산분석(ANOVA)

- 등분산에 대한 ANOVA검정은 두 모집단의 평균이 동일하다(분산의 비율F-값을 이용)는 귀무가설을 검정함
- $H_0$: 여러 집단 간의 평균은 차이가 없다 .
- $H_1$: 적어도 한 그룹의 평균은 차이가 있다.

**예제)**

작업자 A,B,C 중에 일의 능률의 평균의 차이가 존재할까?

## 추론통계

### 분산분석(ANOVA)

- 작업자에 따라서 생산량의 차이가 있는지 비교 분석
  - 요인(Factor) = 작업자
  - 측정값에 영향을 미치는 요인(factor)이 1개인 실험 : 일원분산분석(one-way ANOVA)
  - 일원 분산 분석은 종속 변수의 평균 사이에 유의한 차이가 있는지 알아볼 수 있도록 함

| 작업자 A | 작업자 B | 작업자 C |
|---|---|---|
| 90 | 75 | 82 |
| 85 | 80 | 76 |
| 77 | 71 | 88 |

| Score | 작업자 |
|---|---|
| 90 | A |
| 85 | A |
| 77 | A |
| 75 | B |
| 80 | B |
| 71 | B |
| 82 | C |
| 76 | C |
| 88 | C |

## 분산분석(ANOVA)

- **Between Variance**

| Score | 작업자 | Group Mean | Between Variance |
|---|---|---|---|
| 90 | A | 84 | 12.67 |
| 85 | A | 84 | 12.67 |
| 77 | A | 84 | 12.67 |
| 75 | B | 75.33 | 26.11 |
| 80 | B | 75.33 | 26.11 |
| 71 | B | 75.33 | 26.11 |
| 82 | C | 82 | 2.43 |
| 76 | C | 82 | 2.43 |
| 88 | C | 82 | 2.43 |

$(84-80.44)^2$
$(84-80.44)^2$
$(84-80.44)^2$

$\Rightarrow 3 \times (84-80.44)^2$

$(75.33-80.44)^2$
$(75.33-80.44)^2$
$(75.33-80.44)^2$

$\Rightarrow 3 \times (75.33-80.44)^2$

$(82-80.44)^2$
$(82-80.44)^2$
$(82-80.44)^2$

$\Rightarrow 3 \times (82-80.44)^2$

$$SS_{between} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{total})^2$$

$$SS_{between} = 123.66$$

$$df = (k-1) = 2$$

## 분산분석(ANOVA)

- **Within Variance**

| Score | 작업자 | Group Mean | Within Variance |
|---|---|---|---|
| 90 | A | 84 | |
| 85 | A | 84 | |
| 77 | A | 84 | |
| 75 | B | 75.33 | |
| 80 | B | 75.33 | |
| 71 | B | 75.33 | |
| 82 | C | 82 | |
| 76 | C | 82 | |
| 88 | C | 82 | |

$(90-84)^2$
$(85-84)^2$
$(77-84)^2$

$(75-75.33)^2$
$(80-75.33)^2$
$(71-75.33)^2$

$(82-82)^2$
$(76-82)^2$
$(88-82)^2$

$$SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

$SS_{within}=$

$df = N - k = 9 - 3 = 6$

## 분산분석(ANOVA)

- 측정값에 영향을 미치는 요인(factor)이 1개인 실험 : 일원분산분석(one-way ANOVA)
  - 다중비교
  - $H_0 : \mu_1 = \mu_2 = \mu_3$ → 작업자는 일의 능률에 영향을 미치지 않는다.
  - $H_1$ : not $H_0$ → 어느 집단의 평균에 차이가 발생하고 있는지를 파악하기 위해
    - $\mu_1 \neq \mu_2 = \mu_3$
    - $\mu_1 = \mu_2 \neq \mu_3$
    - $\mu_1 \neq \mu_3 = \mu_2$

## 추론통계

**예제)**

**작업자의 작업 점수에 대한 집단의 평균에 차이가 발생하고 있는지를 파악**
**A: [12, 15, 14, 10]**
**B: [20, 18, 22, 21]**
**C: [25, 23, 27, 26]**

$$SS_{between} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{total})^2$$

$$SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2$$

$$F = \frac{SS_{between}/(k-1)}{SS_{within}/(N-k)}$$

**v1 : between-group(자유도)**
**v2 : within-group(자유도)**

$$\bar{X}_A = 12.75,\ \bar{X}_B = 20.25,\ \bar{X}_c = 25.25$$
$$\bar{X}_{total} = 19.42$$

$$SS_{between} = 4(12.75 - 19.42)^2 + n_B(\bar{X}_B - \bar{X}_{total})^2 + n_c(\bar{X}_c - \bar{X}_{total})^2 = 316.6$$

$SS_{within}(Group\ A)$

$$(12 - 12.75)^2 = -0.75^2 = 0.5625$$
$$(15 - 12.75)^2 = 2.25^2 = 5.0625$$
$$(14 - 12.75)^2 = 1.25^2 = 1.5625$$
$$(10 - 12.75)^2 = -2.75^2 = 7.5625$$

$0.5625+5.0625+1.5625+7.5625=14.7$

$SS_{within}(Group\ A+B+C)$
$14.7+8.75+8.75=32.2$

$$F = \frac{316/(3-1)}{32.2/(12-3)} \approx 44.2$$

## 분산분석(ANOVA)

```
# 1. 데이터 불러오기
data <- read.csv("class_scores.csv", stringsAsFactors = TRUE)

# 2. ANOVA 분석
anova_result <- aov(Score ~ Class, data = data)

# 3. 분석 결과 출력
summary(anova_result)

# 4. 사후검증 출력
library(multcomp) #사후검정

tukey_result <- glht(anova_result, linfct = mcp(Class = "Tukey"))
summary(tukey_result)
```

## 추론통계

### 분산분석(ANOVA)

- 일원 분산분석(one-way ANOVA): 두 개 이상의 수준 또는 범주가 있는 하나의 독립변수가 있고, 이를 종속변수와 비교
- 이원 분산분석(Two-way ANOVA): 두 개의 독립 변수가 존재하며, 이 변수들은 개별적으로, 연관적으로 종속변수에 영향을 미치는지 판단하는 것(ex) 식이요법과 운동 수준이 체중 감량에 어떤 영향을 미치는지 알고 싶을 때

| 종속변수 | 독립변수 |
|---|---|
| 50 | A 학원 |
| 60 | B 학원 |
| 70 | A 학원 |
| 85 | B 학원 |
| 67 | C 학원 |
| 88 | A 학원 |
| 54 | C 학원 |

**일원분산분석**

| 종속변수 | 식이요법 | 운동 수준 |
|---|---|---|
| 50 | A 요법 | 5(명목/순위) |
| 60 | B 요법 | 5(명목/순위) |
| 70 | A 요법 | 6(명목/순위) |
| 85 | B 요법 | 8(명목/순위) |
| 67 | C 요법 | 5(명목/순위) |
| 88 | A 요법 | 8(명목/순위) |
| 54 | C 요법 | 4(명목/순위) |

**이원분산분석**

## 추론통계

### 분산분석(ANOVA)

#### 일원 분산분석

- 세 개 이상의 독립적인 그룹의 평균 간에 통계적으로 유의미한 차이가 있는지 여부를 테스트함
- 기본적으로 데이터의 전체 분산을 "그룹 간" 분산과 "그룹 내" 분산으로 분해한 다음 F-통계량을 둘의 비율로 계산함

#### 이원 분산분석

- 각 개별 요인의 효과를 테스트하는 것뿐만 아니라 종속 변수에 대한 두 요인 사이에 상호 작용이 있는지를 확인하는데 사용됨
- 상호 작용 : 한 요인의 효과가 다른 요인의 수준에 따라 달라지는가 궁금한 것

## 추론통계

### 가설 검정 및 추론통계

- 독립변수 : 가설의 원인이 되는 변수, 종속변수에 영향을 미치는 변수
- 종속변수 : 가설의 결과가 되는 변수, 독립변수로 영향을 받는 변수

| 원인 | 결과 |
| --- | --- |
| 독립 변수(Independent Variable) | 종속 변수(Dependent Variable) |
| 설명 변수(Explanatory Variable) | 반응 변수(Response Variable) |
| 예측 변수(Predictor Variable) | 결과 변수(Outcome Variable) |
| | 표적 변수(Target Variable) |

## 추론통계

### 가설 검정 및 추론통계

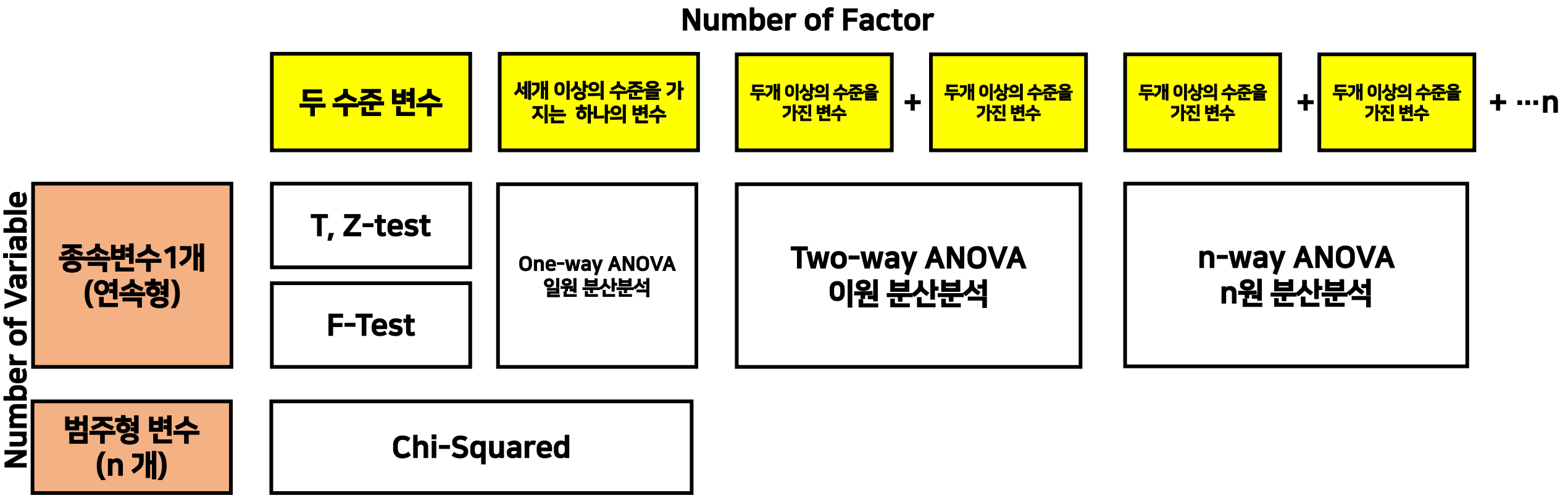
| 원인 | 결과 |
| --- | --- |
| 독립 변수(Independent Variable) | 종속 변수(Dependent Variable) |

- 독립변수: 종속변수에 영향을 미치는 변수 ⟷ 통계적 독립 : 두 변수 간에 아무런 관련성이 없음
- 상호작용 있음 → 두 변수가 개별적으로 효과가 없더라도 결합될 때 특정 결과에 영향을 줄 수 있음
- 통계적 독립이 성립하지 않으며, 상관 가능성 있음 → 두 변수가 같이 있을 때 효과가 달라짐
  - Ex) 수면시간과 카페인 섭취는 개별로는 집중력에 미미한 효과지만, 함께 작용하면 큰 차이를 만들 수 있음

- 상관관계 있음 → 통계적으로 독립이 아님, 반드시 상호작용이 있는 것은 아님
- 통계적 독립이 아님, 상호작용은 아닐 수 있음 → 두 변수는 같이 움직임
  - Ex) 공부 시간이 늘어나면 성적도 올라감

## 추론통계

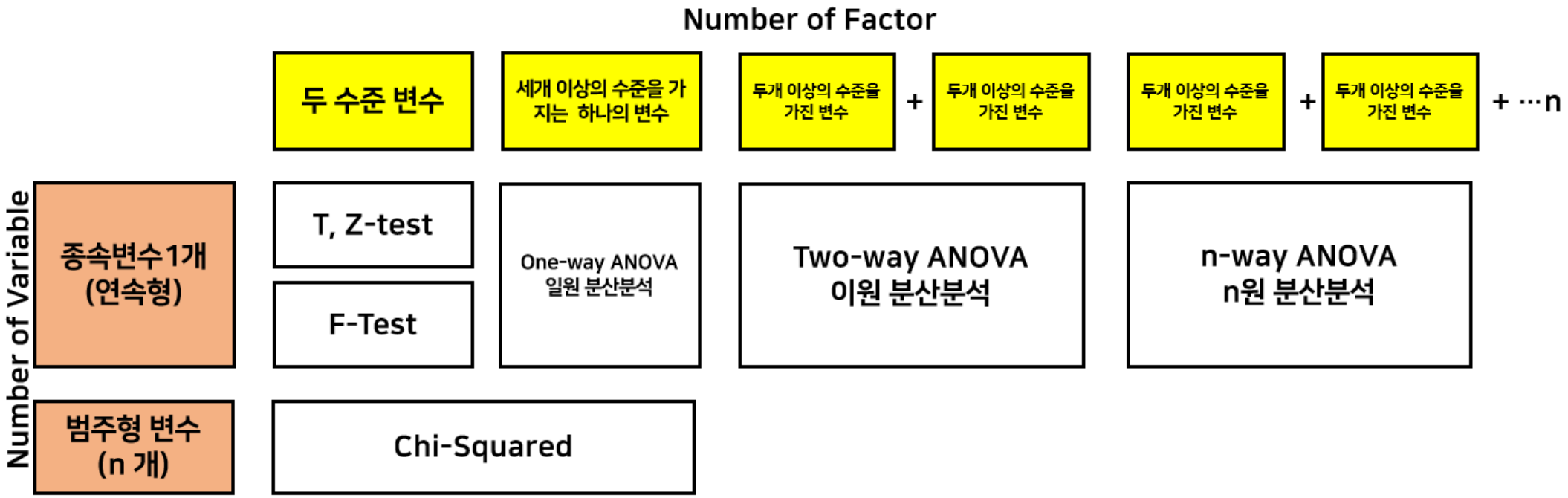
### 가설 검정 및 추론통계



## 추론통계

### 가설 검정 및 추론통계

- 분산분석(ANOVA)의 종속변수는 연속형
- ANOVA는 데이터 그룹 간의 차이가 통계학적으로 유의한지 알아낼 수 있도록 도와 줌
- ANOVA의 목적은 여러 집단 간 평균이 같다고 볼 수 있는지를 검정하는 것



## 분산분석(ANOVA)

- 측정값에 영향을 미치는 요인(factor)이 2개 : 이원분산분석(two-way ANOVA)

| Growth | 비료 | 물 |
|---|---|---|
| 30 | A | Low |
| 35 | A | High |
| 37 | A | Low |
| 28 | A | High |
| 29 | B | Low |
| 32 | B | High |
| 33 | B | Low |
| 36 | C | High |
| 35 | C | Low |
| 37 | C | High |

## 분산분석(ANOVA)

- 측정값에 영향을 미치는 요인(factor)이 2개인 실험 : 이원분산분석(two-way ANOVA)
  - 다중비교
  - $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ (독립변수 $a$)
  - $H_1$ : 독립변수 a의 집단의 평균 중 하나 이상의 평균이 다름 → 독립변수 a가 유의한 영향(종속변수)을 미침
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0$ (독립변수 $b$)
  - $H_1$ : 독립변수 b의 집단의 평균 중 하나 이상의 평균이 다름 → 독립변수 b가 유의한 영향(종속변수)을 미침

  - 상호작용 효과
  - $H_0$ : 독립변수 a와 독립변수 b 사이에는 상호작용이 없다.(종속변수에 영향을 주지 못함)
  - $H_1$ : 독립변수 a와 독립변수 b 사이에는 상호작용이 있다.(종속변수에 영향을 줌)

## 분산분석(ANOVA)

$$SS_A = n_i \sum_{i=1}^{a} (\bar{X}_i - \bar{X}_{total})^2$$

$$SS_B = n_j \sum_{j=1}^{b} (\bar{X}_j - \bar{X}_{total})^2$$

$$SS_{A\times B} = n_{ij} \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{total})^2$$

$$SS_{Total} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{X}_{ijk} - \bar{X}_{total})^2$$

$$SS_{between} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X}_{total})^2$$

$$SS_{within} = SS_{Total} - SS_A - SS_B - SS_{AXB}$$

**요인 A가 종속 변수에 영향을 미치는지 여부**

$$F_A = \frac{SS_A/n_A - 1}{SS_{within}/(n_A \times n_B \times (n-1))}$$

**요인 B가 종속 변수에 영향을 미치는지 여부**

$$F_B = \frac{SS_B/n_B - 1}{SS_{within}/(n_A \times n_B \times (n-1))}$$

**요인 A, B의 상호작용으로 종속변수에 영향**

$$F_{A\times B} = \frac{SS_{A\times B}/((n_A - 1)(n_B - 1))}{SS_{within}/(n_A \times n_B \times (n-1))}$$

## 분산분석(ANOVA)

$$SS_A = n_i \sum_{i=1}^{a} (\bar{X}_i - \bar{X}_{total})^2$$

$$SS_B = n_j \sum_{j=1}^{b} (\bar{X}_j - \bar{X}_{total})^2$$

$$SS_{A\times B} = n_{ij} \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{X}_{ij} - \bar{X}_i - \bar{X}_j + \bar{X}_{total})^2$$

$$SS_{Total} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (\bar{X}_{ijk} - \bar{X}_{total})^2$$

$$SS_{between} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{X}_{total})^2$$

$$SS_{within} = SS_{Total} - SS_A - SS_B - SS_{AXB}$$

$$F_A = \frac{SS_A / n_A - 1}{SS_{within} / (n_A \times n_B \times (n-1)}$$

$$F_B = \frac{SS_B}{n_B - 1}, F_{AXB} = \frac{SS_{AXB}}{n_A X n_B - 2}, MS_{within} = \frac{SS_{within}}{df_{within}}$$

## 분산분석(ANOVA)

```
grow <- read.csv("growth.csv", stringsAsFactors = TRUE) #문자형 변수를 요소로 변환

# 이원분산분석(독립변수들간 상호작용X)
anova_result <- aov(growth ~ fertilizer + water, data = grow)
summary(anova_result)

# 이원분산분석(독립변수들간 상호작용)
anova_result <- aov(growth ~ fertilizer * water, data = grow)
summary(anova_result)
```

## 사후검정

- Tukey 검증 : 각 그룹의 샘플의 수가 같고 분산이 유사할 경우에 주로 사용
- games-howell 검증 : 각 그룹의 샘플의 수가 서로 다르고, 분산이 유사하지 않을 경우에 주로 사용 (불균형 설계)

```
library(car)     #등분산검정
library(rstatix) #games_howell_test

#사후검정(fertilizer)
#p-value<0.05일 때 games-howell 검증사용가능(등분산검정을 통해 등분산이 아닐 경우)
leveneTest(growth ~ fertilizer, data = grow)

tukey_result <- glht(anova_result, linfct = mcp(fertilizer = "Tukey"))
games_howell_test(grow, growth ~ fertilizer)

#사후검정(water)
leveneTest(growth ~ water, data = grow)

tukey_result <- glht(anova_result, linfct = mcp(water = "Tukey"))
games_howell_test(grow, growth ~ water)
```

## 사후검정

- Tukey 검증 : 각 그룹의 샘플의 수가 같고 분산이 유사할 경우에 주로 사용
- games-howell 검증 : 각 그룹의 샘플의 수가 서로 다르고, 분산이 유사하지 않을 경우에 주로 사용 (불균형 설계)

```
# 사후검정(fertilizer*water)
# 교호작용을 하나의 그룹 변수로 통합

grow$group <- interaction(grow$fertilizer, grow$water) #변수 묶어서 하나의 변수로 만들어 줌
group_model <- aov(growth ~ group, data = grow)

leveneTest(growth ~ group, data = grow) #귀무가설 기각 시(p-value<0.05), games-Howell 검증사용 가능

tukey_result <- glht(group_model, linfct = mcp(group = "Tukey"))
summary(tukey_result)

games_howell_test(grow, growth ~ group)
```

## 추론통계

### 사후검정

- Tukey 검증 : 각 그룹의 샘플의 수가 같고 분산이 유사할 경우에 주로 사용
- games-howell 검증 : 각 그룹의 샘플의 수가 서로 다르고, 분산이 유사하지 않을 경우에 주로 사용 (불균형 설계)

```
read.csv("cafe.csv", stringsAsFactors = TRUE)

#분산분석

#사후검증(Tukey)

카페 손님들의 만족도에 좋은 영향을 미치는 각변수별 항목들은 무엇인가?
```

## 추론통계

### 가설 검정 및 추론통계

#### T-test, Z-test

| A집단 | B집단 |
|---|---|
| 연속형 | 연속형 |
| 연속형 | 연속형 |
| 연속형 | 연속형 |
| 연속형 | 연속형 |

#### 카이제곱 분석

| A요인 / B요인 | B요소_1 | B요소_2 |
|---|---|---|
| A요소_1 | 범주형 | 범주형 |
| A요소_2 | 범주형 | 범주형 |

#### F-test

| A집단 | B집단 |
|---|---|
| 연속형 | 연속형 |
| 연속형 | 연속형 |

#### ANOVA(일원분산분석)

| 집단(종속) | B요인(독립) |
|---|---|
| 연속형 | 범주형_요소1 |
| 연속형 | 범주형_요소2 |
| 연속형 | 범주형_요소3 |

#### ANOVA(이원분산분석)

| 집단(종속) | A요인(독립) | B요인(독립) |
|---|---|---|
| 연속형 | 범주형_요소1 | 범주형_요소1 |
| 연속형 | 범주형_요소2 | 범주형_요소2 |
| 연속형 | 범주형_요소3 | 범주형_요소3 |