데이터과학을 위한 R프로그래밍

6주차. R을 이용한 통계분석

이혜선 교수
포항공과대학교 산업경영공학과

목차

6주차. R을 이용한 통계분석

1차시 두 그룹간 평균비교분석

2차시 짝을 이룬 그룹간 평균비교

3차시 분산분석(ANOVA)

6주차

3차시

분산분석(ANOVA)

-Analysis of Variance-

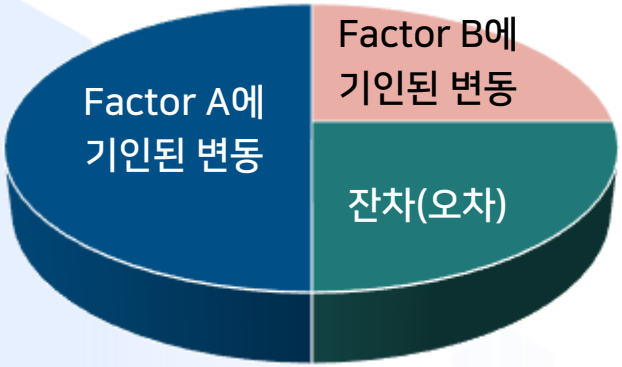
POSTECH

분산분석의 개념

☑ ANOVA (Analysis of Variance) : 전체 분산(variance)을 분할(분석, analysis)하여 어떤 요인(factor)의 영향이 유의한 지(significant)한지 검정하는 방법.

(예) Drug effect (5mg, 10mg, placebo) Age effect(young, old)

factor가 1개일 때의 분산분석모형



전체변동(전체분산)= $\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_{..})^2$

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij},\begin{cases} i=1,2,...,a \\ j=1,2,...,n \end{cases}$$

$\mu=$ an overall mean, $\tau_i = ith$ treatment effect,

$\varepsilon_{ij}=$ experimental error, $N(0,\sigma^2)$

POSTECH

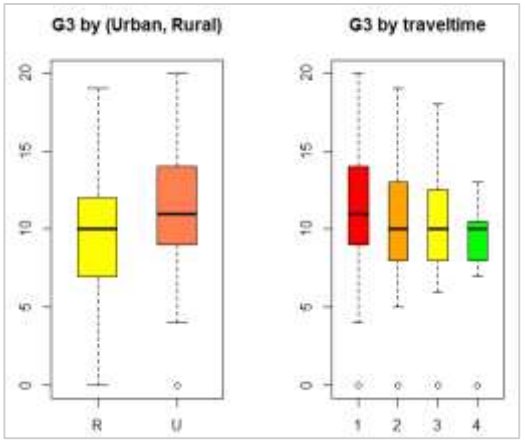
분산분석: Factor가 한 개 일 때

분산분석모형 적용

- (1) 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)
- (2) 통학 시간에 따른 학업성취도 : 통학 시간(factor: 1-4), 학업성적(1-20)

Week5_3에서의 그래프를 이용한 데이터 탐색

```
# 2. boxplot
par(mfrow=c(1,2))
boxplot(G3~address, boxwex = 0.5, col = c("yellow", "coral"), main="G3 by
boxplot(G3~traveltime, boxwex = 0.5, col = c("red","orange","yellow","gree
```



(1) 도심지역 학생들 성적이 외곽지역 학생들보다 높다

(2) 통학시간이 짧은(15분 이내)의 학생들의 성적이 더 높다

POSTECH

분산분석: Factor가 한 개 일 때

- 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)
  - 가설1: 거주지역(R/U)에 따라 G3에 유의한 영향이 있나?
  - aov (타겟변수~factor)

```
# 1. ANOVA by address
a1 <- aov(G3~address)
summary(a1)
```

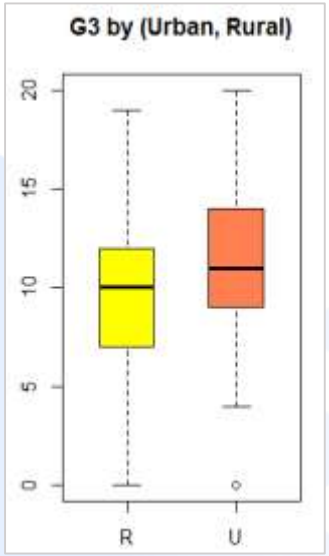
```
> a1 <- aov(G3~address)
> summary(a1)
             Df Sum Sq Mean Sq F value Pr(>F)
address       1     92   92.49   4.445 0.0356 *
Residuals   393   8177   20.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  - p-value=0.035. 유의 수준(α=0.05)에서 0.05보다 작으므로
    ⇒ 거주지역에 따른 학업성적에는 유의한 차이가 있다고 할 수 있음

POSTECH

분산분석: Factor가 한 개 일 때

- 거주지역에 따른 학업성취도 : 거주지역(factor: R/U), 학업성적(1-20)
  - 분산분석 결과는 상자그림으로 본 거주지역에 따른 G3의 차이가 통계적으로 유의하다는 것을 보여줌!!




```
# tapply - give FUN value by address
round(tapply(G3, address, mean),2)
```

```
> round(tapply(G3, address, mean),2)
    R     U
 9.51 10.67
```

G3(Rural)=9.51, G3(Urban)=10.67

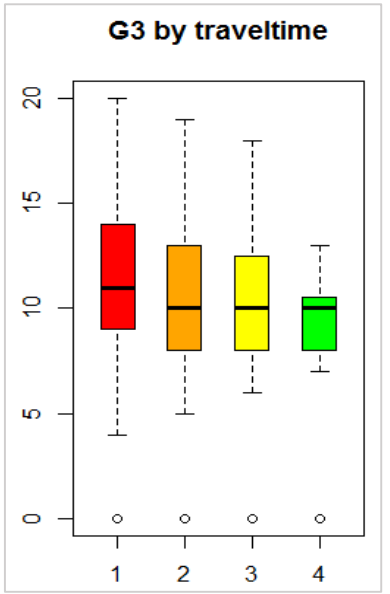
POSTECH

분산분석: Factor가 한 개 일 때

- 통학 시간에 따른 학업성취도 : 통학 시간(factor: 1-4), 학업성적(1-20)
  - 가설2: 통학 시간에 따라 G3에는 유의한 차이가 있나?
  - aov(타겟변수~factor)

```
# 2. ANOVA by traveltime
traveltime<-as.factor(traveltime)
a2 <- aov(G3~traveltime)
summary(a2)
```

```
> # 2. ANOVA by traveltime
> a2 <- aov(G3~traveltime)
> summary(a2)
             Df Sum Sq Mean Sq F value Pr(>F)
traveltime    3    115   38.37    1.84  0.139
Residuals   391   8155   20.86
```



- p-value=0.139, 유의 수준을 0.05로 잡을 때 0.05보다 크므로
  ⇒ 유의수준 0.05에서는 통학 시간에 따른 학업성적에는 유의한 차이가 없다고 할 수 있다
- 그러나 p-value가 0.139이므로 어느정도 차이가 존재함을 알 수 있다.

Traveltime : 통학시간 1(15분이하),2(15-30분),3(30-1시간),4(1시간이상)

POSTECH

분산분석: Factor가 한 개 일 때

☑ 사후검정 : ANOVA에서 어떤 factor의 유의성이 검정되면, 그 다음 단계에 하는 검정

Tukey's Honest Significant Difference Test

```
# should be foctor for Tukey's Honest Sig
TukeyHSD(a2, "traveltime", ordered=TRUE)
plot(TukeyHSD(a2, "traveltime"))
```

μ_1-μ_2 =0이라는 의미는 1그룹과 2그룹간 차이가 없다는 의미

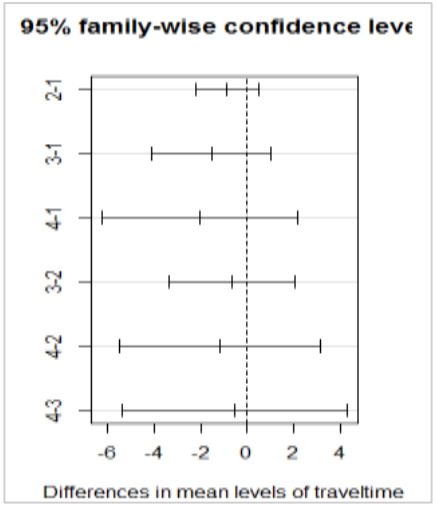
즉 (μ_1-μ_2)의 신뢰 구간에 0이 있다는 것은 차이가 없다고 할 수 있다

```
> TukeyHSD(a2, "traveltime", ordered=TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = G3 ~ traveltime)

$traveltime
         diff        lwr      upr     p adj
3-4 0.5108696 -4.3256048 5.347344 0.9929165
2-4 1.1565421 -3.1623166 5.475401 0.9005404
1-4 2.0321012 -2.1981712 6.262374 0.6021367
2-3 0.6456725 -2.0624782 3.353823 0.9272302
1-3 1.5212316 -1.0432848 4.085748 0.4202138
1-2 0.8755591 -0.4800959 2.231214 0.3429618
```

95% 신뢰구간의 lower bound, upper bound



모든 pairwise 신뢰구간에 0이 포함됨
⇒ 유의한 차이가 없음

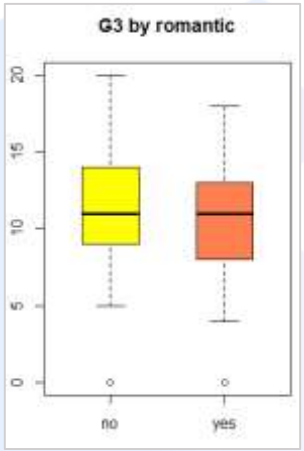
POSTECH

추가 예제: 분산분석

☑ 연애경험여부에 따른 학업성취도 : 연애경험(yes, no), 학업성적(1-20)

```
# 4. ANOVA by romantic
a4 <- aov(G3~romantic)
summary(a4)
# tapply - give FUN value by address
round(tapply(G3,romantic, mean),2)
```

```
> summary(a4)
             Df Sum Sq Mean Sq F value  Pr(>F)
romantic      1    140   139.70   6.753 0.00971 **
Residuals   393   8130    20.69
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
> # tapply - give FUN value by address
> round(tapply(G3,romantic, mean),2)
   no   yes
10.84  9.58
```



연애경험이 있는 경우 학업성적이 유의하게 낮음 (p-value=0.0097)

median은 비슷해 보이지만, 평균은 10.84-9.58=1.26 차이 있음