제 9 강

연구조사방법론

2021. 04. 29

공지사항

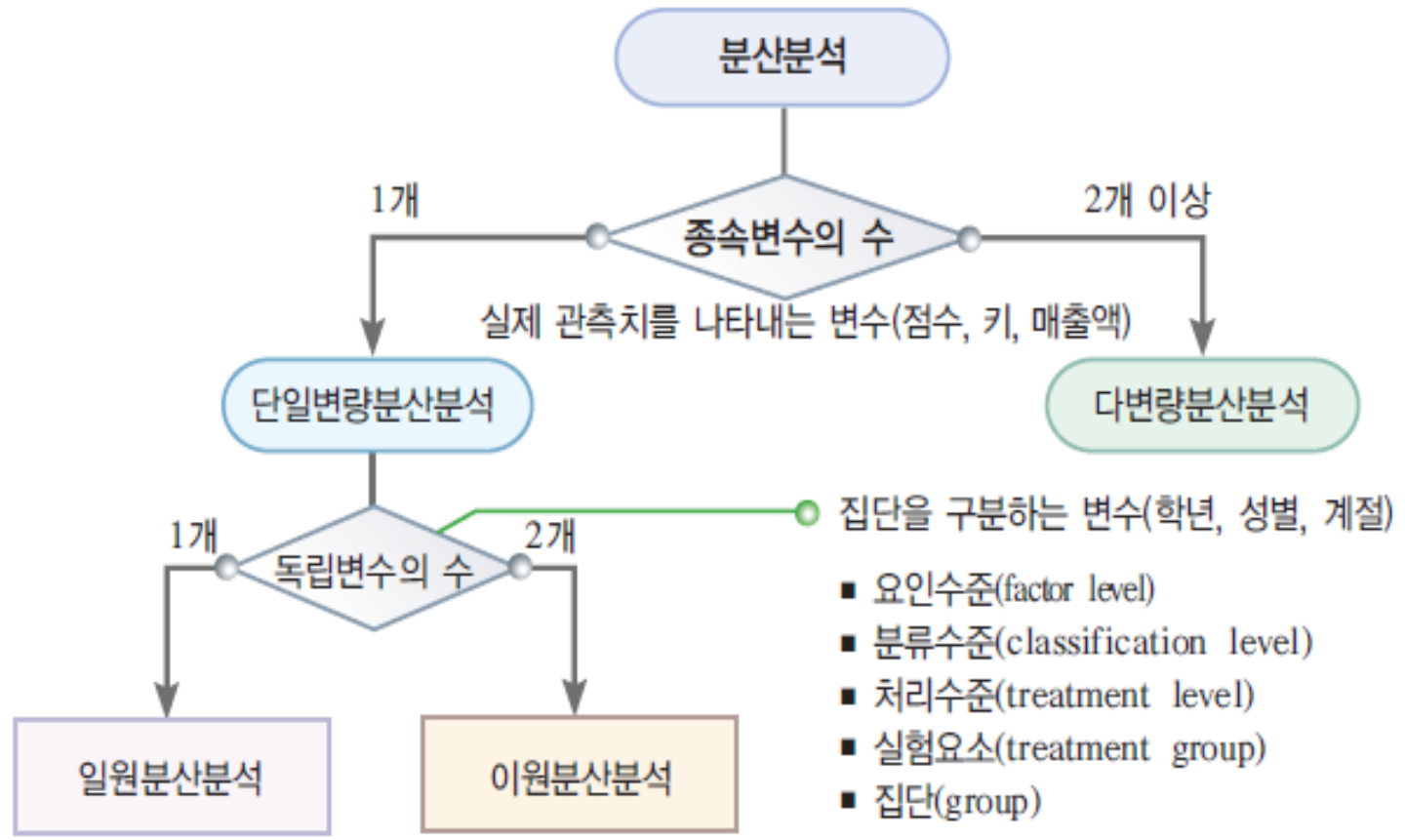
- 7강 까지의 출석과제 마감
- 8강, 9강 출석은 9강 출석과제 제출여부로 처리
- 매주 강의 시간을 이용해서 팀(개별) 질의 응답 시간을 갖고 있습니다.

오늘의 강의 내용

- 공분산분석 - 누구나 15장
- 분산분석II (이원분산분석) - 누구나 14장
- R 실습

복습-분산분석의 종류

- 3개 이상의 집단 간 평균이 서로 다른 지를 검정하는 분석 방법
- 또 다른 각도로는, 독립변수가 종속변수에 미치는 영향을 분석하는 방법 중 하나
  - 종속변수는 연속변수이어야 하며, 독립변수로 구분되는 각각의 집단에 속한 관측치(종속변수 값)의 평균이 통계적으로 유의하게 차이가 있는지를 분석

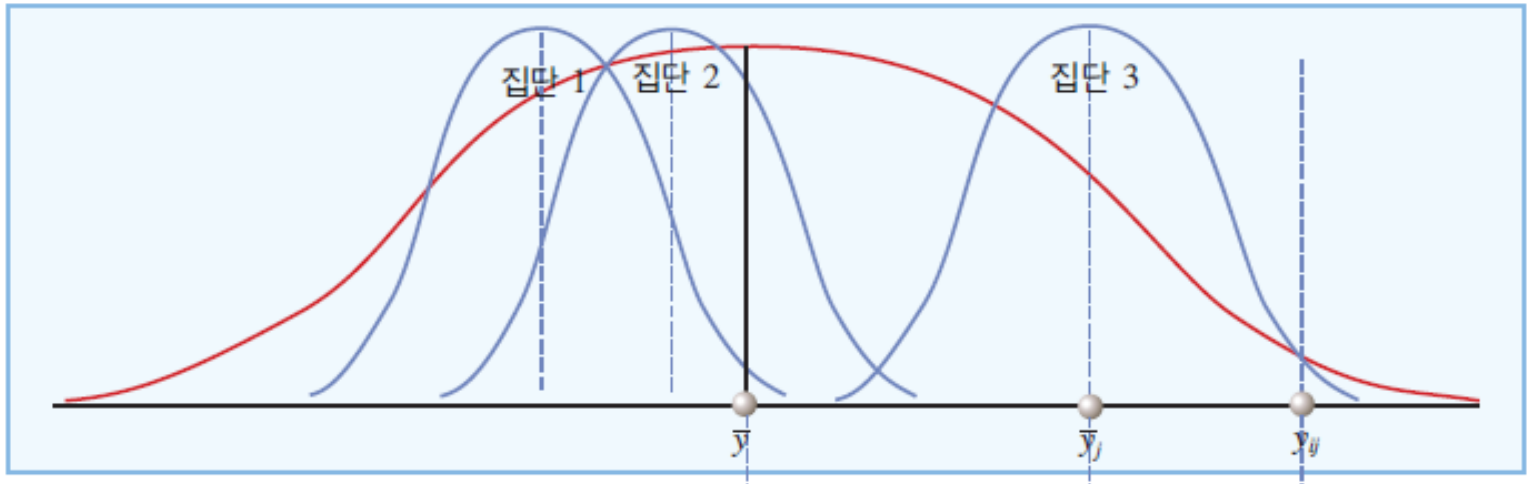


〈그림11-4〉

복습-일원분산분석(One Way ANOVA)

일원분산분석의 개념

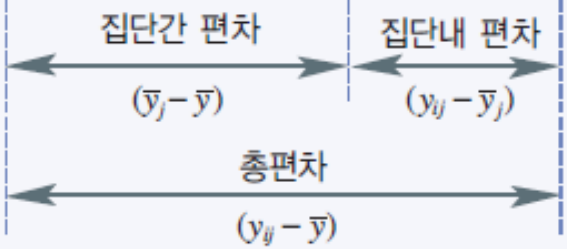
- 집단을 구분하는 독립변수가 1개인 경우 집단간 종속변수의 평균이 서로 다른지를 분석하는 방법



$$y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$$

관측치 = 전체평균 + 집단 j의 처리효과 + 오차

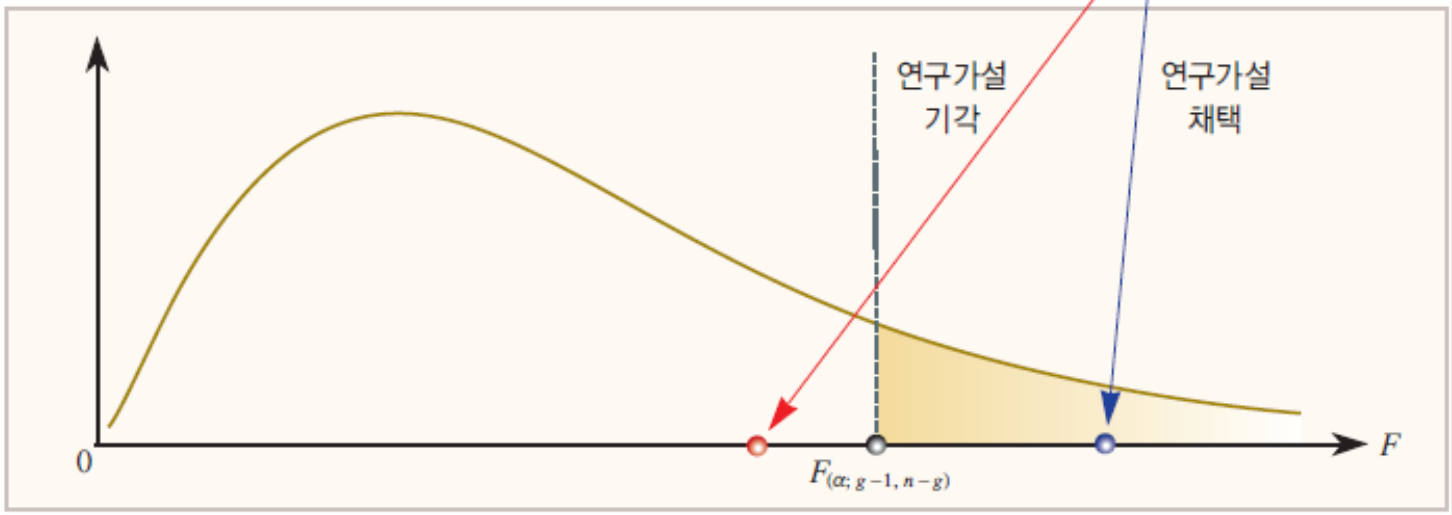
i : 특정한 집단에 속한 관측치를 지정하는 첨자
j : 특정한 집단을 지정하는 첨자



〈그림11-5〉

복습-일원분산분석표와 유의성 검정

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

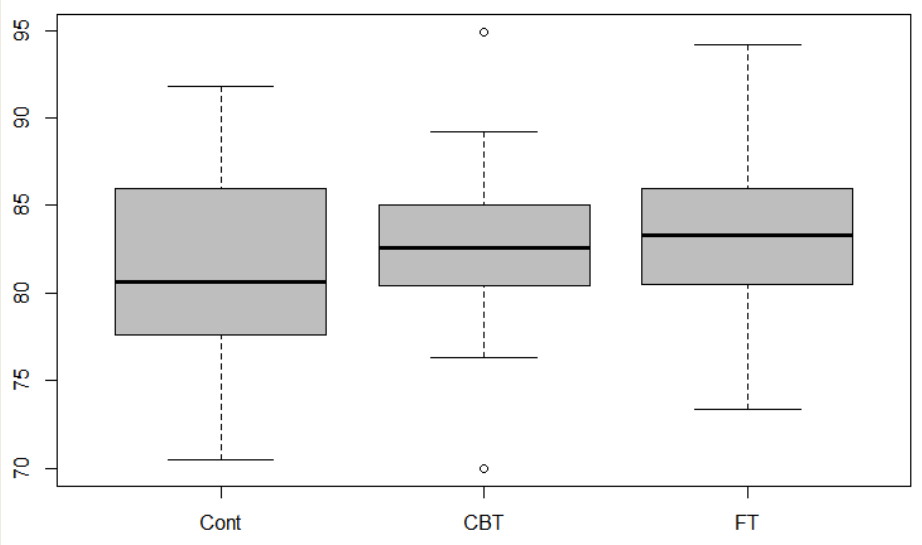


- 분산분석은 집단간 평균제곱(MSB)을 집단내 평균제곱(MSW)으로 나눈 통계량 F값을 검정통계량값으로 하여 집단간 평균의 차이가 통계적으로 유의한지를 분석함
- 연구가설 : 비교하려는 집단들의 평균이 모두 같지는 않음
  - 적어도 한 집단의 평균은 나머지와 차이가 있음

〈그림11-9, 10〉

일원분산분석 연습문제 12.7

- anorexia.csv의 세 그룹 간의 Prewt 평균의 차이가 있는지 일원분석을 해보자



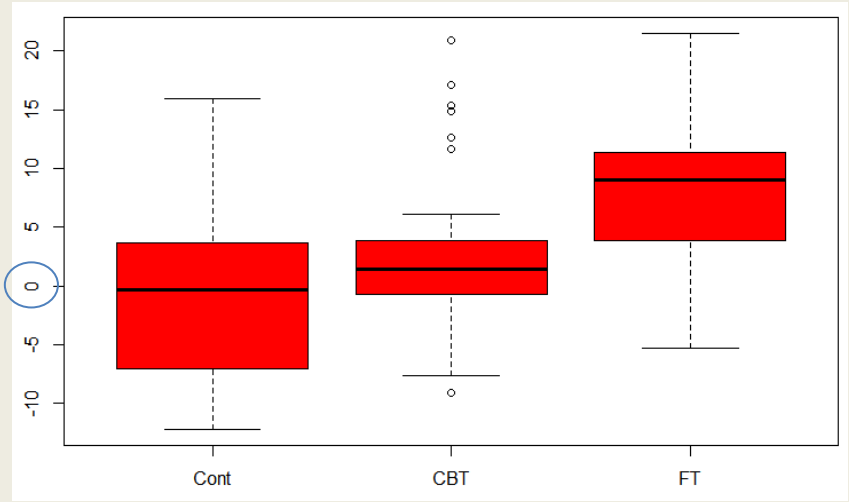

```
> out <- lm(Prewt~Treat, data=anorexia)
> anova(out)
Analysis of Variance Table

Response: Prewt
          Df  Sum Sq Mean Sq F value Pr(>F)
Treat      2   32.57  16.285  0.5995 0.5519
Residuals 69 1874.35  27.164
> shapiro.test(resid(out))  # 잔차의 정규성 검정

        Shapiro-Wilk normality test

data:  resid(out)
W = 0.99241, p-value = 0.9461
```

```
# 치료전 몸무게의 일원분산분석
anorexia <- read.csv("anorexia.csv")
anorexia$Treat = relevel(anorexia$Treat, ref = "Cont")
levels(anorexia$Treat)
boxplot(Prewt~Treat, data=anorexia, col ='grey')

out <- lm(Prewt~Treat, data=anorexia)
anova(out)
shapiro.test(resid(out))  # 잔차의 정규성 검정
```

- 일원분석결과 : 세 그룹 간의 치료전 몸무게 차이가 없다고 분석되었다.
- 정규성 검정결과 잔차의 정규성이 확인되었다.

일원분산분석 연습문제 12.7

- anorexia.csv의 세 그룹간의 전후 몸무게 차이가 있는지 일원분산분석을 한 후 사후검정을 해보자



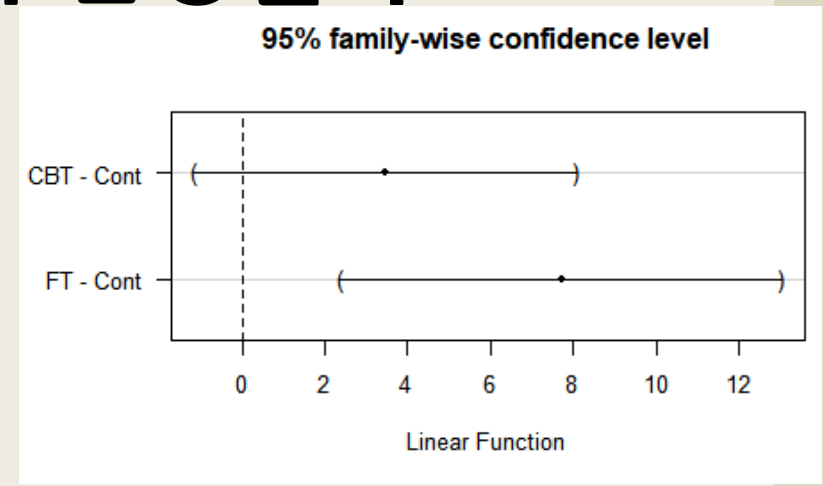

```
# 치료전후 몸무게 차이에 대한 일원분산분석
anorexia$diff <- anorexia$Postwt-anorexia$Prewt
boxplot(diff~Treat, data=anorexia, col ='red')

out <- lm(diff~Treat, data = anorexia)
anova(out)
shapiro.test(resid(out))

# 치료전후 몸무게 차이에 대한 일원분산분석 후 사후검정
install.packages("multcomp")
library(multcomp)

out <- lm(diff~Treat, data = anorexia)
dunnett <- glht(out, linfct = mcp(Treat ="Dunnett"))
summary(dunnett)
plot(dunnett)

tukey <- glht(out, linfct = mcp(Treat ="Tukey"))
summary(tukey)
plot(tukey)
```

```
> anova(out)
Analysis of Variance Table

Response: diff
          Df Sum Sq Mean Sq F value   Pr(>F)
Treat      2  614.6 307.322  5.4223 0.006499 **
Residuals 69 3910.7  56.677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> shapiro.test(resid(out))

        Shapiro-Wilk normality test

data:  resid(out)
W = 0.96723, p-value = 0.05726
```

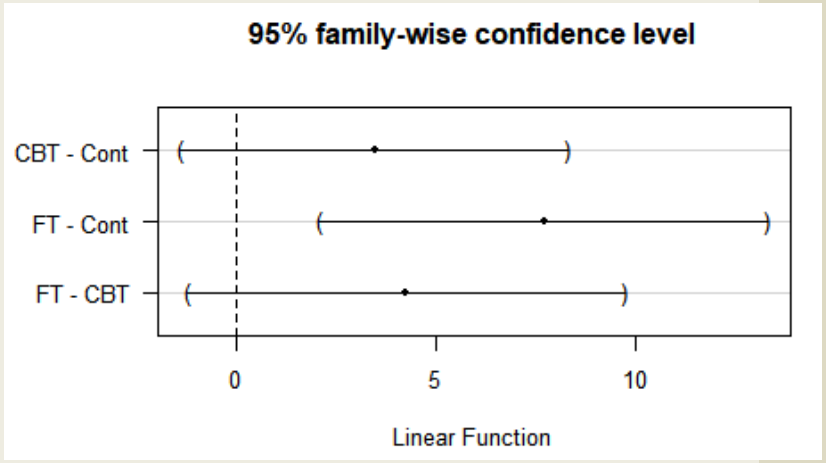
- 일원분석결과 : 세 그룹의 치료전후 몸무게 차이의 평균이 같지 않은 것으로 분석되었다.
- 정규성 검정결과 잔차의 정규성이 확인되었다.

일원분산분석
연습문제 12.7 - 사후검정결과

```
> summary(dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = diff ~ Treat, data = anorexia)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
CBT - Cont == 0    3.457      2.033   1.700  0.16654
FT - Cont == 0     7.715      2.348   3.285  0.00313 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```




```
> summary(tukey)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = diff ~ Treat, data = anorexia)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
CBT - Cont == 0    3.457      2.033   1.700  0.21161
FT - Cont == 0     7.715      2.348   3.285  0.00443 **
FT - CBT == 0      4.258      2.300   1.852  0.16005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```



- 그룹 간 비교결과 FT와 Cont 그룹의 몸무게 전후 차이가 있어서 분산분석에서 세 그룹간 평균의 차이가 있다는 결과를 얻게 된 것으로 분석되었다.

공분산분석(ANCOVA, Analysis of Covariance)

- 분산분석에 연속형 변수(공변량)를 추가한 것
- 궁극적인 목적은 분산분석과 동일하나, 통제가 안 되는 연속형 변수를 추가하여 오차를 줄이고 검정력을 높이는 것이 차이점
- anorexia.csv 예제에서 공변량(Covariate)으로 Prewt(이전 몸무게)을 추가하면 설명이 안 되는 Error가 줄어듦. Treat의 설명 비율이 설명 안되는 Error에 비해 상대적으로 커져서 Treat가 유의하게 나올 가능성이 커짐
  - 일원분산분석
    Diff = Treat + Error
  - 공분산분석
    Diff = Prewt +Treat + Error
    Postwt = Prewt +Treat + Error

종속변수를 Postwt으로 해도
p-value는 같음
Prewt의 p-value는 공분산분석에서
관심대상이 아님

- **R-script**

```
anorexia <- read.csv("anorexia.csv")
anorexia

levels(anorexia$Treat)
anorexia$Treat <- relevel(anorexia$Treat, ref = 'Cont')
levels(anorexia$Treat)
anorexia$Treat <- factor(anorexia$Treat, levels=c('Cont', 'CBT', 'FT'))


out <- lm(Postwt~Prewt+Treat, data=anorexia)
anova(out)
summary(out)
```

공분산분석 R 분석 예

```
> out <- lm(Postwt~Prewt+Treat, data=anorexia)
> anova(out)
Analysis of Variance Table

Response: Postwt
          Df Sum Sq Mean Sq F value    Pr(>F)
Prewt      1  506.5  506.51 10.4017 0.0019364 **
Treat      2  766.3  383.14  7.8681 0.0008438 ***
Residuals 68 3311.3   48.70
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p-value를 봤을 때, 세 그룹간 몸무게 변화의 평균에 통계적으로 유의한 차이가 있는 것으로 분석됨

```
> summary(out)

Call:
lm(formula = Postwt ~ Prewt + Treat, data = anorexia)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1083  -4.2773  -0.5484   5.4838  15.2922

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.6740    13.2167   3.456 0.000950 ***
Prewt         0.4345     0.1612   2.695 0.008850 **
TreatCBT      4.0971     1.8935   2.164 0.033999 *
TreatFT       8.6601     2.1931   3.949 0.000189 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.978 on 68 degrees of freedom
Multiple R-squared:  0.2777,    Adjusted R-squared:  0.2458
F-statistic: 8.713 on 3 and 68 DF,  p-value: 5.719e-05
```

- Reference 그룹인 Cont와 비교한 TreatCBT와 TreatFT의 평균 차이에 대한 p-value를 봤을 때, 유의한 차이가 있는 것으로 분석됨
- 그룹이 3개 이상이므로 사후 검정으로 p-value를 계산(2개의 그룹 비교에서 자유도에 차이가 나며 t분포의 형태가 바뀌면 임계치가 다름)

공분산분석 R 분석 예

공분산분석의 사후검정

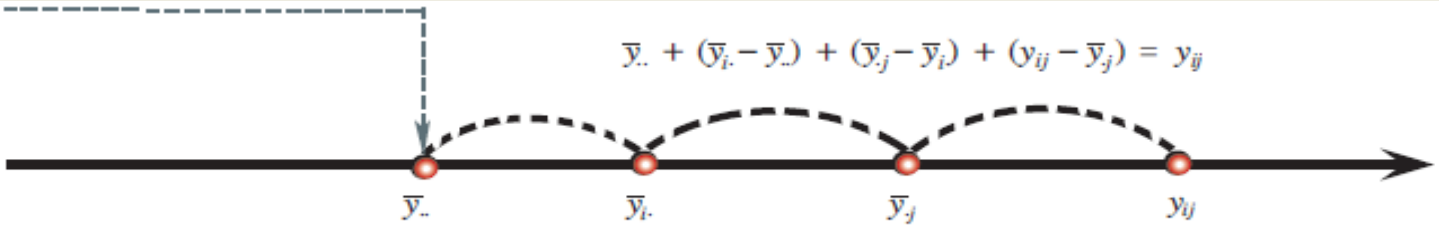
```
> summary(dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = Postwt ~ Prewt + Treat, data = anorexia)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
CBT - Cont == 0     4.097      1.893   2.164 0.062939 .
FT - Cont == 0      8.660      2.193   3.949 0.000373 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

- 사후검정으로 다중비교 한 결과 CBT 그룹은 Cont 그룹과 유의한 차이가 있다고 볼 수 없다고 나옴

```
> summary(dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = diff ~ Treat, data = anorexia)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
CBT - Cont == 0     3.457      2.033   1.700  0.16654
FT - Cont == 0      7.715      2.348   3.285  0.00313 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

- **분산분석의 다중비교 결과와 비교하여 보면 치료효과의 차이가 더 크게 계산되는 것을 확인할 수 있음**

이원분산분석-주효과 검정만 가능한 경우



$$
\begin{aligned}
y_{ij} &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{i.}) + (y_{ij} - \bar{y}_{.j}) \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..} + \bar{y}_{..} - \bar{y}_{i.}) + (y_{ij} - \bar{y}_{.j}) \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + \left[(\bar{y}_{..} - \bar{y}_{i.}) + (y_{ij} - \bar{y}_{.j})\right] \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{..} - \bar{y}_{i.} - \bar{y}_{.j} + y_{ij})
\end{aligned}
$$

여기서, y_{ij} : 요인 i와 요인 j에 의해 구분되는 집단에 속한 관측치

$\bar{y}_{..}$: 요인을 구분하지 않는 전체평균

$\bar{y}_{i.}$: 요인 i에 의해서만 구분되는 집단의 평균

$\bar{y}_{.j}$: 요인 j에 의해서만 구분되는 집단의 평균

i : 요인 i에 의해서 구분되는 c개의 집단을 구분하는 첨자

j : 요인 j에 의해서 구분되는 g개의 집단을 구분하는 첨자

- 관측치(y_{ij}) = 전체평균($\bar{y}_{..}$)
  + 전체평균에서부터 관측치가 속한 요인 i의 평균까지의 거리($\bar{y}_{i.} - \bar{y}_{..}$)
  + 관측치가 속한 요인 i의 평균에서부터 관측치가 속한 요인 j의 평균까지의 거리($\bar{y}_{.j} - \bar{y}_{i.}$)
  + 관측치가 속한 요인 j의 평균에서부터 관측치까지의 거리($y_{ij} - \bar{y}_{.j}$)
- 일원분산분석에서와 같이 2개 요인의 독립적인 효과를 추정하기 위해 식을 변형하면 위와 같은 결과를 얻을 수 있음

〈그림11-18〉

이원분산분석 주효과 검정만 가능한 경우 - 편차, 제곱합, 자유도, 평균제곱

한 집단

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{..} - \bar{y}_{i.} - \bar{y}_{.j} + y_{ij})$$

(총편차) = (요인 i에 의한 편차) + (요인 j에 의한 편차) + (요인 i와 j에 의하여 설명할 수 없는 편차)
(요인 i의 주효과) (요인 j의 주효과) (오차)

여러 개 집단

$$\sum_{i=1}^{c}\sum_{j=1}^{g}(y_{ij}-\bar{y}_{..})^2 = \sum_{j}^{g}\sum_{i=1}^{c}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i}^{c}\sum_{j=1}^{g}(\bar{y}_{.j}-\bar{y}_{..})^2 + \sum_{i=1}^{c}\sum_{j=1}^{g}(\bar{y}_{..}-\bar{y}_{i.}-\bar{y}_{.j}+y_{ij})^2$$

(총제곱합) (요인 i에 의한 제곱합) (요인 j에 의한 제곱합) (오차제곱합)

$$SST = SSB_i + SSB_j + SSE$$

$$cg-1 = c-1 + g-1 + (c-1)(g-1)$$

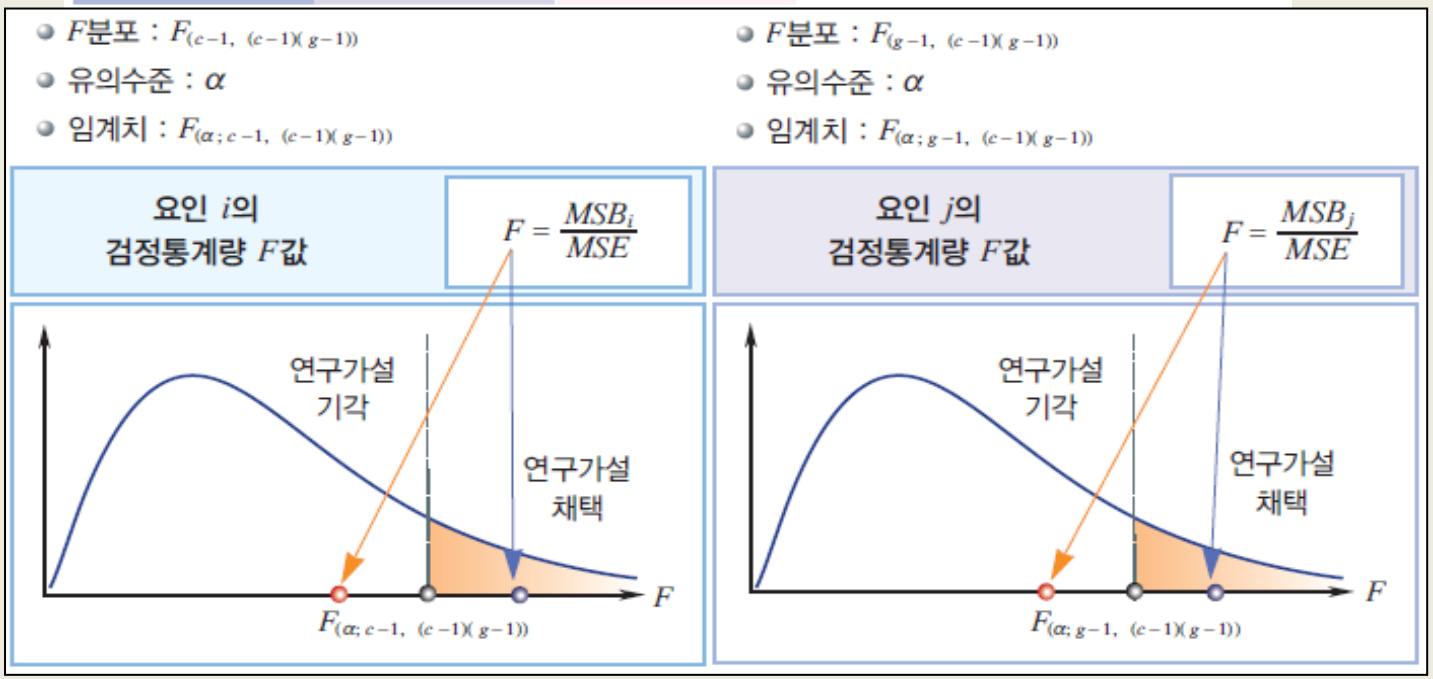
총제곱합의 자유도 = 요인 i에 의한 제곱합의 자유도 + 요인 j에 의한 제곱합의 자유도 + 오차 제곱합의 자유도

여기서, 요인 i에 의해 구분되는 집단의 수: c개
요인 j에 의해 구분되는 집단의 수: g개

〈그림11-19, 20, 21〉

이원분산분석표와 유의성 검정 -주효과분석만 가능한 경우

원 천	제곱합(SS)	자유도(df)	평균제곱(MS)	F비
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	$cg-1$		



각 요인에 대한 가설검정을 별도로 함

〈그림11-22, 23〉

이원분산분석 사례-Excel
반복이 없는 이원 배치법, 주효과 분석만

- 어느 문구판매전문회사에서 임금지급방식에 대해 고민하고 있다. (1) 고정급만 받는 방식, (2) 고정급과 성과급을 함께 받는 방식, (3) 성과급만 받는 방식 세 가지 이다.
- 임금지급방식에 따라 판매실적이 달라질 것이라고 생각하고 있는데 통계적으로 유의한 지 판단하고자 하였다.
- 또한, 영업사원의 경험이 판매실적에 영향을 미칠 것이라고 판단하여 근무년수에 따라 8개 그룹으로 나누었다.

- 자료 : 6개월간 24명의 판매실적(각 대안별 8명)

Group1	Fixed	FixPIncen	Incentive
1	4500	4430	5810
2	4580	4740	5420
3	4200	4530	4800
4	4860	4830	5100
5	5040	5100	5460
6	4740	4920	6180
7	4320	4140	4680
8	4410	4320	4620

분산 분석: 반복 없는 이원 배치법

요약표	관측수	합	평균	분산
1	3	14740	4913.333	604233.3
2	3	14740	4913.333	198933.3
3	3	13530	4510	90300
4	3	14790	4930	21900
5	3	15600	5200	51600
6	3	15840	5280	615600
7	3	13140	4380	75600
8	3	13350	4450	23700
Fixed	8	36650	4581.25	80412.5
FixPIncen	8	37010	4626.25	106169.6
Incentive	8	42070	5258.75	313955.4

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
인자 A(행)	2436263	7	348037.5	4.564426	0.007625	2.764199
인자 B(열)	2296233	2	1148117	15.05727	0.000324	3.738892
잔차	1067500	14	76250			
계	5799996	23				

이원분산분석 사례-R 분석
반복이 없는 이원 배치법, 주효과 분석만

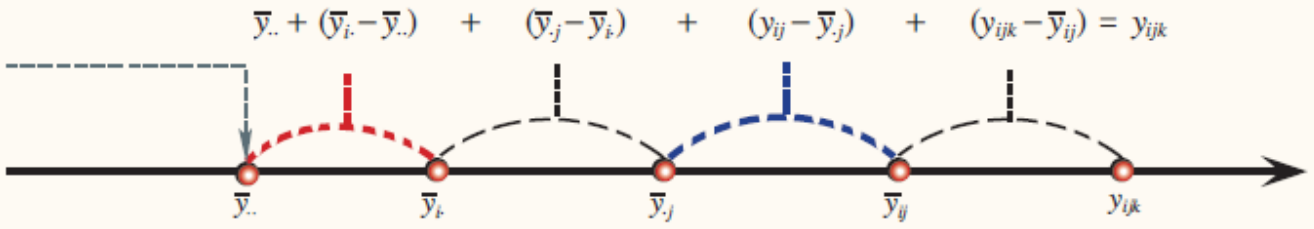
Group	payType	SalesRec
1	F	4500
2	F	4580
3	F	4200
4	F	4860
5	F	5040
6	F	4740
7	F	4320
8	F	4410
1	FI	4430
2	FI	4740
3	FI	4530
4	FI	4830
5	FI	5100
6	FI	4920
7	FI	4140
8	FI	4320
1	I	5810
2	I	5420
3	I	4800
4	I	5100
5	I	5460
6	I	6180
7	I	4680
8	I	4620

```
#import data file 2

library(readxl)
salesRecord2 <- read_excel("salesRecord2.xlsx")
View(salesRecord2)

# encode a vector as a factorr
salesRecord2$PT<- factor(salesRecord2$payType)
salesRecord2$GR<- factor(salesRecord2$Group)

# lm(), anova(), aov()
gc.out2 <- lm(SalesRec ~ GR + PT, data = salesRecord2)
anova(gc.out2)

aov.out2 <- aov(SalesRec ~ GR + PT, data = salesRecord2)
summary(aov.out2)
```

```
> # lm(), anova(), aov()
> gc.out2 <- lm(SalesRec ~ GR + PT, data = salesRecord2)
> anova(gc.out2)
Analysis of Variance Table

Response: SalesRec
          Df  Sum Sq Mean Sq F value    Pr(>F)
GR         7 2436262  348038  4.5644 0.0076252 **
PT         2 2296233 1148117 15.0573 0.0003242 ***
Residuals 14 1067500   76250
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> aov.out2 <- aov(SalesRec ~ GR + PT, data = salesRecord2)
> summary(aov.out2)
            Df  Sum Sq Mean Sq F value   Pr(>F)
GR           7 2436262  348038   4.564 0.007625 **
PT           2 2296233 1148117  15.057 0.000324 ***
Residuals   14 1067500   76250
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

이원분산분석
-주효과, 상호작용효과 검정 가능

● 주효과와 상호작용효과 검정이 가능한 이원분산분석에서의 관측치



$$
\begin{aligned}
y_{ijk} &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{i.}) + (\bar{y}_{ij} - \bar{y}_{.j}) + (y_{ijk} - \bar{y}_{ij}) \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} \boxed{-\bar{y}_{..} + \bar{y}_{..}} - \bar{y}_{i.}) + (\bar{y}_{ij} - \bar{y}_{.j}) + (y_{ijk} - \bar{y}_{ij}) \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + \left[(\bar{y}_{..} - \bar{y}_{i.}) + (y_{ij} - \bar{y}_{.j})\right] + (y_{ijk} - \bar{y}_{ij}) \\
&= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{..} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{ij}) + (y_{ijk} - \bar{y}_{ij})
\end{aligned}
$$

- 관측치(y_{ijk})는 전체 평균값에 전체 평균에서부터 관측치가 속한 요인 i에 의해 구분되는 집단의 평균($\bar{y}_{i.}$)까지의 거리, 요인 i에 의해 구분되는 집단의 평균($\bar{y}_{i.}$)에서부터 요인 j에 의해 구분되는 집단의 평균($\bar{y}_{.j}$)까지의 거리, 요인 j에 의해 구분되는 집단의 평균($\bar{y}_{.j}$)에서부터 요인 i와 요인 j를 모두 사용하여 구분되는 집단의 평균($\bar{y}_{ij}$)까지의 거리, 그리고 요인 i와 요인 j를 모두 사용하여 구분되는 집단의 평균($\bar{y}_{ij}$)에서부터 관측치(y_{ijk})까지의 거리를 모두 더한 값이 됨
- 일원분산분석에서와 같이 2개 요인의 독립적인 효과를 추정하기 위해 식을 변형하면 편차로 이루어진 식을 구할 수 있음

〈그림11-25〉

이원분산분석, 주효과, 상호작용효과 검정 가능 - 편차, 제곱합, 자유도, 평균제곱

한 집단

$$(y_{ijk}-\bar{y}_{..}) = (\bar{y}_{i.}-\bar{y}_{..}) + (\bar{y}_{.j}-\bar{y}_{..}) + (\bar{y}_{..}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{ij}) + (y_{ijk}-\bar{y}_{ij})$$

(총편차) = (요인 i에 의한 편차) (요인 i의 주효과) + (요인 j에 의한 편차) (요인 j의 주효과) + (요인 i와 요인 j의 상호작용에 의한 편차) (요인 i와 요인 j의 상호작용효과) + (요인 i와 요인 j에 의해 설명되지 않는 집단내 편차) (오차)

여러 개 집단

$$\sum_{i=1}^{c}\sum_{j=1}^{g}\sum_{k=1}^{h_{ij}}(y_{ijk}-\bar{y}_{..})^2 = \sum_{i}^{c}\sum_{j}^{g}\sum_{k=1}^{h_{ij}}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i}^{c}\sum_{j}^{g}\sum_{k=1}^{h_{ij}}(\bar{y}_{.j}-\bar{y}_{..})^2 + \sum_{i}^{c}\sum_{j}^{g}\sum_{k=1}^{h_{ij}}(\bar{y}_{..}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{ij})^2 + \sum_{i}^{c}\sum_{j}^{g}\sum_{k=1}^{h_{ij}}(y_{ijk}-\bar{y}_{ij})^2$$

(총제곱합) = (요인 i에 의한 제곱합) + (요인 j에 의한 제곱합) + (요인 i와 j에 의한 제곱합) + (오차 제곱합)

$$SST = SSB_i + SSB_j + SSB_{ij} + SSE$$

$\sum_i^c\sum_j^g h_{ij}-1$	$c-1$	$g-1$	$(c-1)(g-1)$	$\sum_i^c\sum_j^g (h_{ij}-1)$
SST	SSB_i	SSB_j	SSB_{ij}	SSE
총제곱합의 자유도	요인 i에 의한 집단간 제곱합의 자유도	요인 j에 의한 집단간 제곱합의 자유도	요인 i와 요인 j에 의한 집단간 제곱합의 자유도	오차 제곱합의 자유도

여기서, c : 요인 i에 의해 구분되는 집단의 수(첨자 i로 구분함)
g : 요인 j에 의해 구분되는 집단의 수(첨자 j로 구분함)
h : 요인 i와 요인 j에 의해 구분되는 집단 내의 관측치 수

〈그림11-26, 27, 28〉

이원분산분석표와 유의성 검정
-주효과, 상호작용효과 분석 가능한 경우

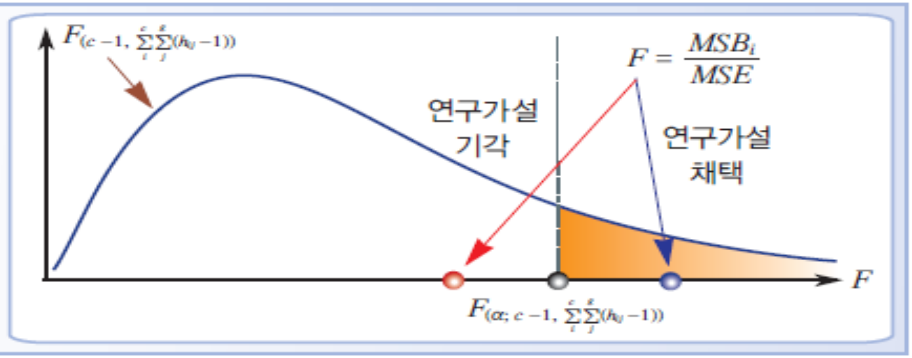
원 천	제곱합(SS)	자유도(df)	평균제곱(MS)	F비
요인 i	SSB_i	$c-1$	MSB_i	$\frac{MSB_i}{MSE}$
요인 j	SSB_j	$g-1$	MSB_j	$\frac{MSB_j}{MSE}$
요인 i와 j의 상호작용	SSB_{ij}	$(c-1)(g-1)$	MSB_{ij}	$\frac{MSB_{ij}}{MSE}$
오차	SSE	$\sum_{i}^{c}\sum_{j}^{g}(h_{ij}-1)$	MSE	여기서, c : 요인 i에 의해 구분되는 집단의 수 g : 요인 j에 의해 구분되는 집단의 수 h : 요인 i와 요인 j로 구분되는 집단의 관측치 수
총(합계)	SST	$\sum_{i}^{c}\sum_{j}^{g}h_{ij}-1$		

〈그림11-29〉

이원분산분석표와 유의성 검정
-주효과, 상호작용효과 분석 가능한 경우

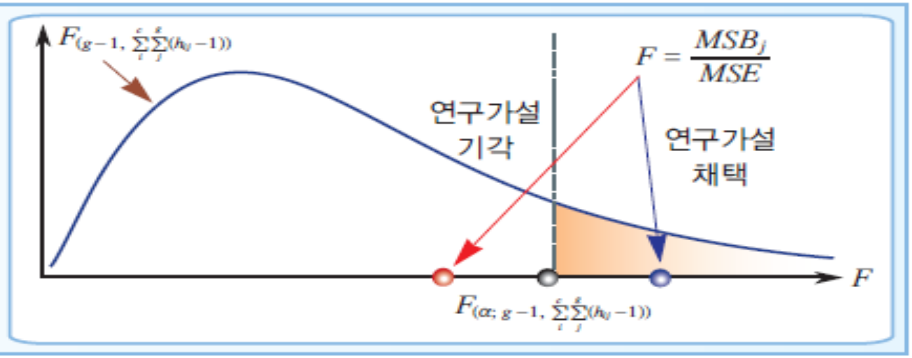
요인 i의 주효과 검정

- 분포 : $F_{(c-1,\ \sum_i\sum_j(h_{ij}-1))}$
- 유의수준 : α
- 임계치 $F_{(\alpha;\ c-1,\ \sum_i\sum_j(h_{ij}-1))}$
- 검정통계량 F값 $F = \dfrac{MSB_i}{MSE}$



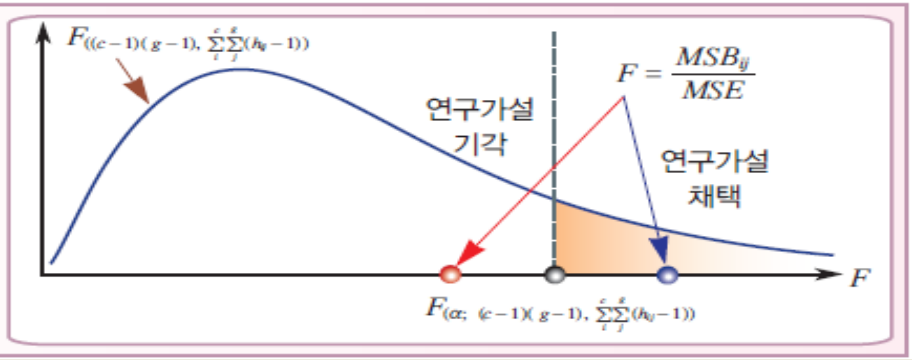
요인 j의 주효과 검정

- 분포 : $F_{(g-1,\ \sum_i\sum_j(h_{ij}-1))}$
- 유의수준 : α
- 임계치 $F_{(\alpha;\ g-1,\ \sum_i\sum_j(h_{ij}-1))}$
- 검정통계량 F값 $F = \dfrac{MSB_j}{MSE}$



상호작용효과 검정 요인 i와 j의

- 분포 : $F_{((c-1)(g-1),\ \sum_i\sum_j(h_{ij}-1))}$
- 유의수준 : α
- 임계치 $F_{(\alpha;\ (c-1)(g-1),\ \sum_i\sum_j(h_{ij}-1))}$
- 검정통계량 F값 $F = \dfrac{MSB_{ij}}{MSE}$



각 요인과 상호작용 효과에 대한 가설검정을 별도로 함

〈그림11-30〉

이원분산분석 사례-Excel
반복이 있는 이원 배치법, 주효과, 상호효과 분석

- 어느 문구판매전문회사에서 임금지급방식에 대해 고민하고 있다. (1) 고정급만 받는 방식, (2) 고정급과 성과급을 함께 받는 방식, (3) 성과급만 받는 방식 세 가지 이다.
- 임금지급방식에 따라 판매실적이 달라질 것이라고 생각하고 있는데 통계적으로 유의한 지 판단하고자 하였다.
- 또한, 판매지역이 판매실적에 영향을 미칠 것이라고 판단하여 네 지역의 자료를 수집하였다.

- 자료 : 6개월간 24명의 판매실적 (임금지급 대안과 지역별 2개 관측치)

판매지역	Fixed	FixPIncen	Incentive
서울	4500	4430	5810
서울	4580	4740	5420
부산	4200	4530	4800
부산	4860	4830	5100
대구	5040	5100	5460
대구	4740	4920	6180
광주	4320	4140	4680
광주	4410	4320	4620

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
인자 A(행)	2154713	3	718237.5	11.08462	0.000896	3.490295
인자 B(열)	2296233	2	1148117	17.71899	0.000262	3.885294
교호작용	571500	6	95250	1.470002	0.26801	2.99612
잔차	777550	12	64795.83			
계	5799996	23				

이원분산분석 사례-R 분석
반복이 있는 이원 배치법, 주효과, 상호효과 분석

Group	payType	SalesRec
Seoul	F	4500
Seoul	F	4580
Busan	F	4200
Busan	F	4860
Daegu	F	5040
Daegu	F	4740
Gwangju	F	4320
Gwangju	F	4410
Seoul	FI	4430
Seoul	FI	4740
Busan	FI	4530
Busan	FI	4830
Daegu	FI	5100
Daegu	FI	4920
Gwangju	FI	4140
Gwangju	FI	4320
Seoul	I	5810
Seoul	I	5420
Busan	I	4800
Busan	I	5100
Daegu	I	5460
Daegu	I	6180
Gwangju	I	4680
Gwangju	I	4620

```r
#import data file 3

library(readxl)
salesRecord3 <- read_excel("salesRecord3.xlsx")
View(salesRecord3)

# encode a vector as a factorr
salesRecord3$PT<- factor(salesRecord3$payType)
salesRecord3$LOC<- factor(salesRecord3$City)

# lm(), anova(), aov()
gc.out3 <- lm(SalesRec ~ LOC + PT + LOC*PT, data = salesRecord3)
anova(gc.out3)

aov.out3 <- aov(SalesRec ~ LOC + PT + LOC*PT, data = salesRecord3)
summary(aov.out3)
```

```
> gc.out3 <- lm(SalesRec ~ LOC + PT + LOC*PT, data = salesRecord3)
> anova(gc.out3)
Analysis of Variance Table

Response: SalesRec
          Df  Sum Sq Mean Sq F value   Pr(>F)
LOC        3 2154713  718238  11.085 0.000896 ***
PT         2 2296233 1148117  17.719 0.000262 ***
LOC:PT     6  571500   95250   1.470 0.268010
Residuals 12  777550   64796
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> aov.out3 <- aov(SalesRec ~ LOC + PT + LOC*PT, data = salesRecord3)
> summary(aov.out3)
             Df  Sum Sq Mean Sq F value   Pr(>F)
LOC           3 2154713  718238   11.09 0.000896 ***
PT            2 2296233 1148117   17.72 0.000262 ***
LOC:PT        6  571500   95250    1.47 0.268010
Residuals    12  777550   64796
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

다변량 분산분석의 개념

- 다수의 종속변수에서 독립변수 집단 간의 차이가 있는지 검증하는 방법
  - 종속변수끼리 상관관계가 있을 때 사용하는 모델로 ANOVA에서 밝힐 수 없는 결합된 평균의 차이를 밝힐 수 있음
    * 종속변수가 서로 독립이면 종속변수 개수만큼 ANOVA분석 시행
  - 연구가설 예 : 광고유무에 따라 광고효과(선호도, 태도, 구매의도)가 있다.
- MANOVA의 가정
  - 관측치가 서로 독립적
  - 각 집단의 분산과 공분산 행렬이 동일(Levene's test, Box's M test)
  - 모든 종속변수들은 다변량 정규분포를 따름
- 다변량 분산분석의 검정값
  - Pillai's Trace : 양의 값, 값이 크면 처리효과가 모형에 미치는 영향이 크다는 의미
  - Wilks's Lambda : 0과 1사이의 값, 값이 작으면 처리효과가 모형에 미치는 영향이 크다는 의미
  - Hotelling's Trace : 양의 값, 값이 크면 처리효과가 모형에 미치는 영향이 크다는 의미
  - Roy's Largest Root : 양의 값, 값이 크면 처리효과가 모형에 미치는 영향이 크다는 의미
- 단순 대비 검정 : 어떤 그룹에서 차이가 났는지 분석
- 개체 간 효과 검정 : 각 종속변수별로 ANOVA 분석

다변량 분산분석 예제

광고경험	선호도	태도	구매의도
아니오	4	4	4
예	7	7	7
아니오	4	6	4
예	4	5	6
아니오	4	6	6
아니오	5	5	5
아니오	4	5	4
아니오	7	4	7
예	3	5	2
예	4	1	4
아니오	5	2	2
아니오	4	6	3
예	7	7	7
.	.	.	.
.	.	.	.

- 종속변수 : 광고효과(선호도, 태도, 구매의도)
- 독립변수 : 광고경험
- 가설검정
  - 귀무가설 : 광고경험 유무에 따라 광고효과는 같을 것이다.
  - 대립가설 : 광고경험 유무에 따라 광고효과는 다를 것이다.
- 절차
  - 상관분석
  - 일반선형모형
  - 다변량 분산분석

다변량 분산분석 예제 R활용 분석

```
AD <- read.csv("manova.csv")
Y = cbind(AD$선호도, AD$태도, AD$구매의도)  #종속변수 구성
cor(Y)    #종속변수간 상관성 분석

out_manova <- manova(Y~AD$광고경험)     # MANOVA

summary(out_manova, test=c("Pillai"))
summary(out_manova, test=c("Wilks"))
summary(out_manova, test=c("Hotelling-Lawley"))
summary(out_manova, test=c("Roy"))

summary.aov(out_manova)    # 개체간 효과분석
```

```
> cor(Y)
          [,1]      [,2]      [,3]
[1,] 1.0000000 0.6473402 0.8735925
[2,] 0.6473402 1.0000000 0.6739773
[3,] 0.8735925 0.6739773 1.0000000
```

➔ 상관성이 있음

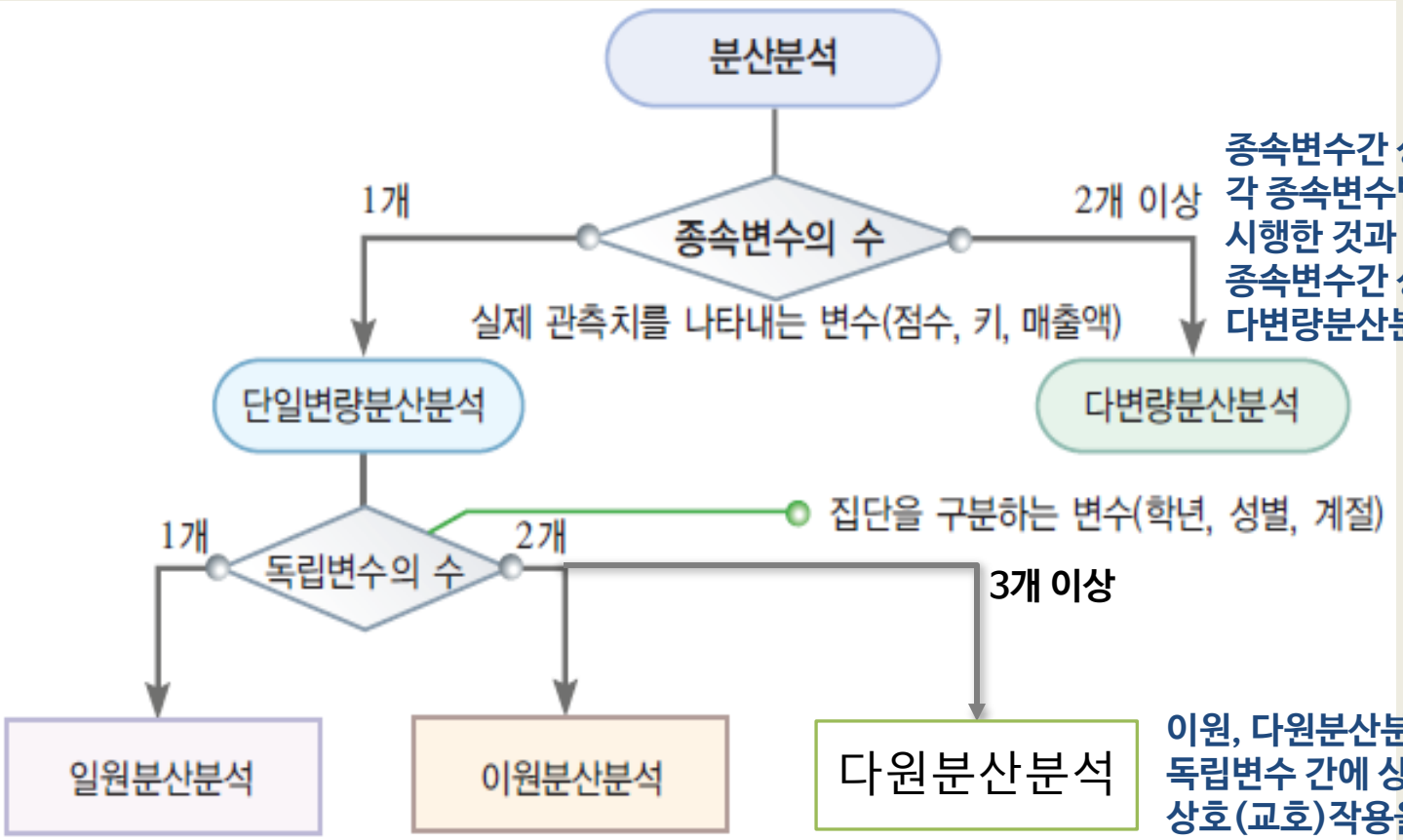
```
> summary.aov(out_manova)
 Response 1 :
             Df  Sum Sq Mean Sq F value Pr(>F)
AD$광고경험   1   0.007 0.00724  0.0035 0.9526
Residuals    88 179.593 2.04083

 Response 2 :
             Df  Sum Sq Mean Sq F value Pr(>F)
AD$광고경험   1   1.208  1.2079   0.575 0.4503
Residuals    88 184.848  2.1005

 Response 3 :
             Df  Sum Sq Mean Sq F value Pr(>F)
AD$광고경험   1   0.985 0.98542  0.4198 0.5187
Residuals    88 206.570 2.34739
```

```
> summary(out_manova, test=c("Pillai"))
             Df   Pillai approx F num Df den Df Pr(>F)
AD$광고경험   1 0.041323   1.2356      3     86 0.3018
Residuals    88
> summary(out_manova, test=c("Wilks"))
             Df   Wilks approx F num Df den Df Pr(>F)
AD$광고경험   1 0.95868   1.2356      3     86 0.3018
Residuals    88
> summary(out_manova, test=c("Hotelling-Lawley"))
             Df Hotelling-Lawley approx F num Df den Df Pr(>F)
AD$광고경험   1          0.043104   1.2356      3     86 0.3018
Residuals    88
> summary(out_manova, test=c("Roy"))
             Df       Roy approx F num Df den Df Pr(>F)
AD$광고경험   1 0.043104   1.2356      3     86 0.3018
Residuals    88
```

➔ 귀무가설을 기각하지 못함: 광고경험 유무에 따라 광고효과에 차이가 없다.

분산분석의 종류 정리

- 3개 이상의 집단 간 평균이 서로 다른 지를 검정하는 분석 방법
- 또 다른 각도로는, 독립변수가 종속변수에 미치는 영향을 분석하는 방법 중 하나
  - 종속변수는 연속변수이어야 하며, 독립변수로 구분되는 각각의 집단에 속한 관측치(종속변수 값)의 평균이 통계적으로 유의하게 차이가 있는지를 분석



종속변수간 상호작용이 없는 경우
각 종속변수별로 단일변량분산분석을
시행한 것과 같음.
종속변수간 상호작용이 있는 경우
다변량분산분석을 시행

이원, 다원분산분석에서는
독립변수 간에 상호작용이 있을 때
상호(교호)작용을 포함하여 분석함

분산분석 R 실습

- 실습순서
  - 누구나 Chapter 15
  - Manova
  - 누구나 Chapter 14

warpbreaks.csv의 데이터

- 종속변수 : breaks(break발생 횟수)
- 독립변수 : wool(털실의 종류), tension(실의 장력)
  - wool : A, B
  - tension : L, M, H

```
> tapply(breaks, wool, mean)
       A        B
31.03704 25.25926
> tapply(breaks, tension, mean)
       L        M        H
36.38889 26.38889 21.66667
> tapply(breaks, list(wool, tension), mean)
         L        M        H
A 44.55556 24.00000 24.55556
B 28.22222 28.77778 18.77778
```

- 두 변수의 어떤 조합이 가장 적은 또는 많은 Break를 발생시키는 지 조사하고 싶음

통계분석 R 실습

- 연습문제 14.4
  - warpbreaks.csv를 이용하여
    - 상호작용(interaction)을 무시하고 **wool과 tension 만으로 이원분산분석 후**
    - **tension을 Dunnett의 다중비교 방법으로 비교하여라.**
    - **회귀진단을 실시하고 잔차의 정규성을 검정하여 통계분석의 적절성을 나타내어라.**
  - 분석결과를 문서로 간략하게 정리해보자
    - 가설설정, R-script, R 분석결과 및 그래프, 결과 해석, 결론
    - 정리된 워드 파일만 9강 출석과제에 업로드