



제 8 강



연구조사방법론

2021. 4. 22





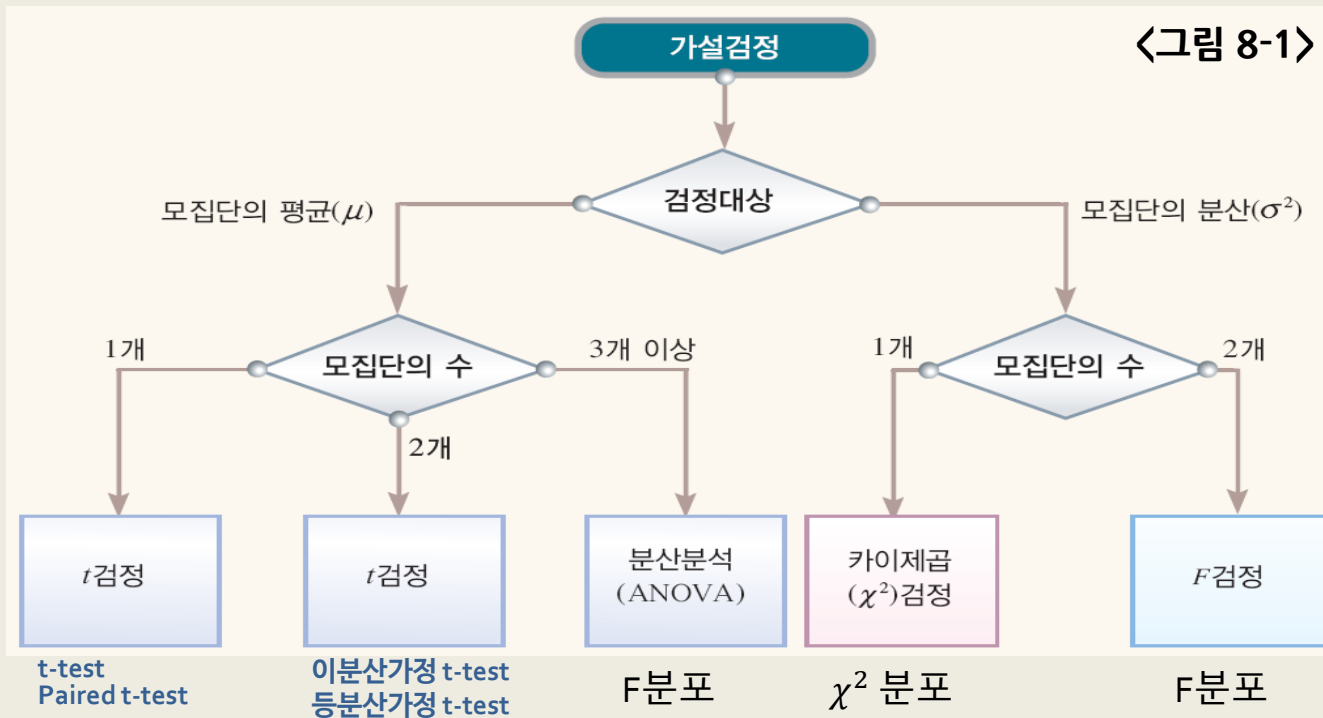
오늘의 강의 내용

- ❑ 실습문제 풀이
 - ▣ 누구나 5.7
 - ▣ 누구나 7.8
 - ▣ 연구조사방법론 수강생 대상 설문조사 분석
- ❑ 분석방법의 선택(독립변수, 종속변수 구분이 있을 때)
- ❑ 분산분석 I (일원분산분석)



지난주 복습 - 평균과 분산에 대한 가설검정 (비모수 검정 추가)

❑ 모집단의 평균과 분산에 대한 가설검정



비모수검정 →

Wilcoxon
signed-rank test

Wilcoxon
rank-sum test

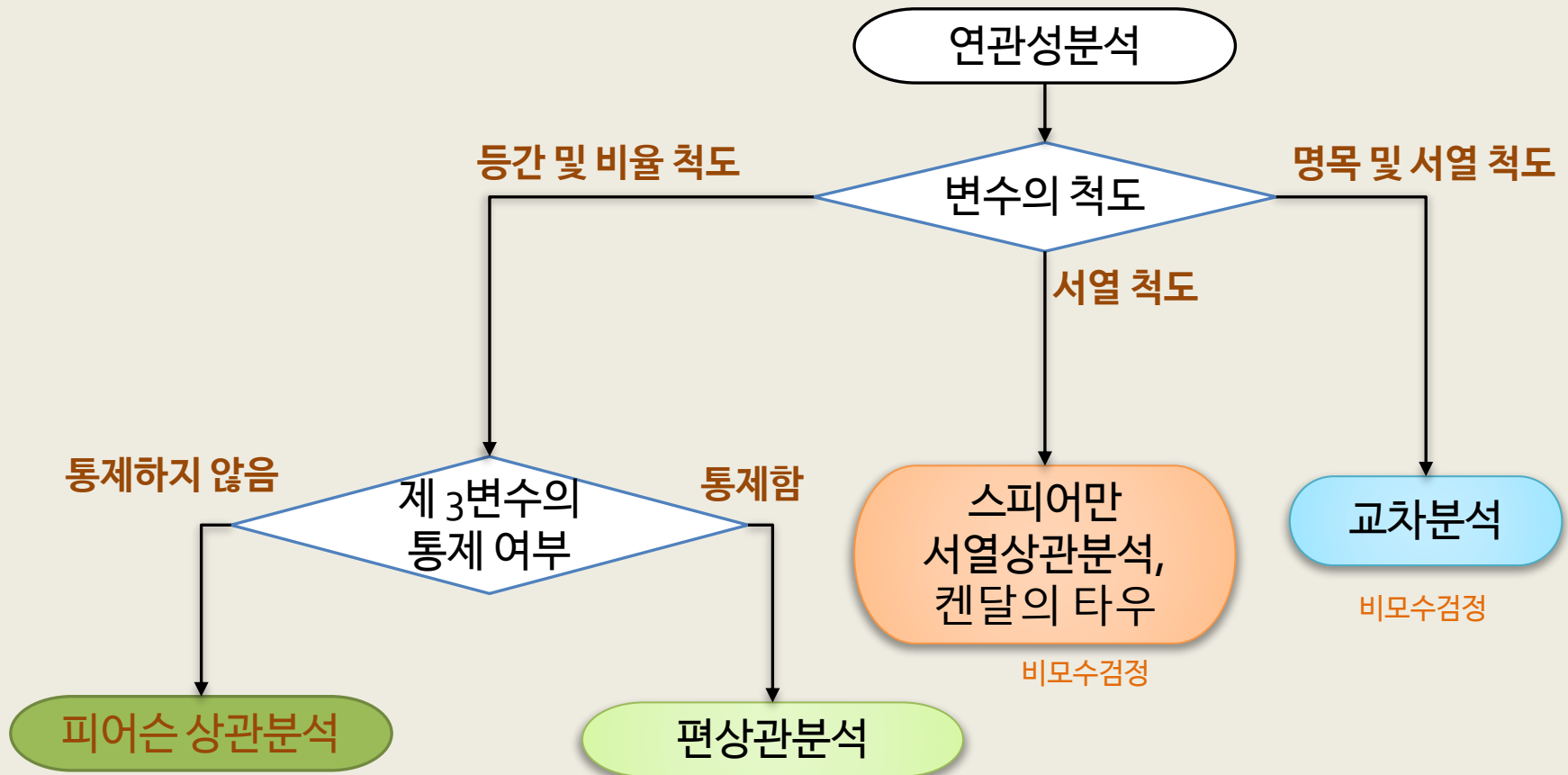
Kruskal-Wallis test

표본데이터의 분포가 정규성을 만족하지 않는 경우
샤피로 윌크 검정 `shapiro.test(x)`



지난주 복습- 연관성분석

- 2개 변수들 사이의 연관성을 파악하는 방법



<그림10-1>



연습문제 5.7 풀이

- ❑ anorexia (거식증)의 Treat (치료방법)이 “Cont”인 관찰치만을 사용하여 다음을 실행하여 보자
 - ▣ shapiro.test()로 정규성 검정을 해보자
 - ▣ t.test()로 paired t-test를 해보자

▶ R Script

```
anorexia <- read.csv("anorexia.csv")  
with(anorexia[anorexia$Treat=="Cont", ], shapiro.test(Postwt-Prewt))  
with(anorexia[anorexia$Treat=="Cont", ], t.test(Postwt-Prewt))
```



연습문제 5.7 풀이

□ shapiro.test() 정규성 검정 결과

shapiro-wilk normality test

```
data: Postwt - Prewt  
W = 0.95189, p-value = 0.2567
```

귀무가설 H_0 : 데이터는 정규분포를 따른다
대립가설 H_1 : 데이터는 정규분포를 따르지 않는다.

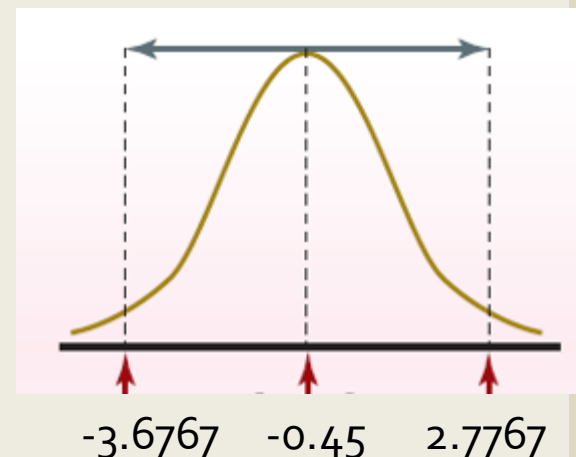
→ p-Value $0.2567 < 0.05$ 가 아니므로 귀무가설 채택.
데이터는 정규분포를 따른다고 볼 수 있다.

▶ Paired t-test 결과 (추정과 가설검증을 위한 분석결과를 제공)

One sample t-test

```
data: Postwt - Prewt  
t = -0.28723, df = 25, p-value = 0.7763  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
-3.676708 2.776708  
sample estimates:  
mean of x  
-0.45
```

▶ 추정(모수 추정)이라면

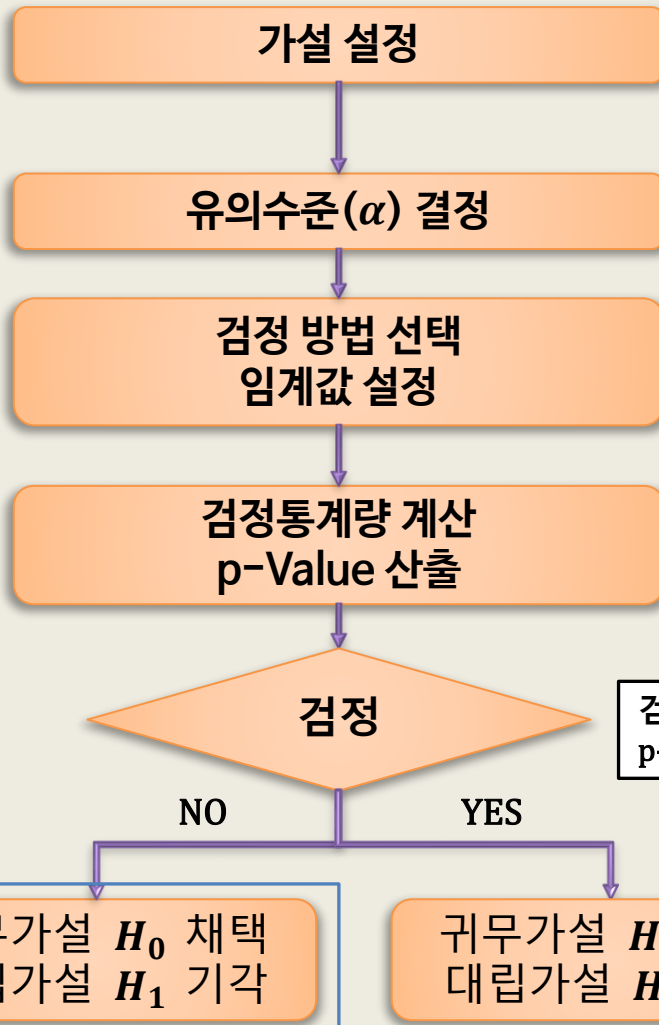




연습문제 5.7 풀이

모집단에 대해 알려진 가설(귀무가설)에 대한 검정. 표본의 통계량으로 모집단에 대한 가설을 적합한지 판단

▶ 가설검정

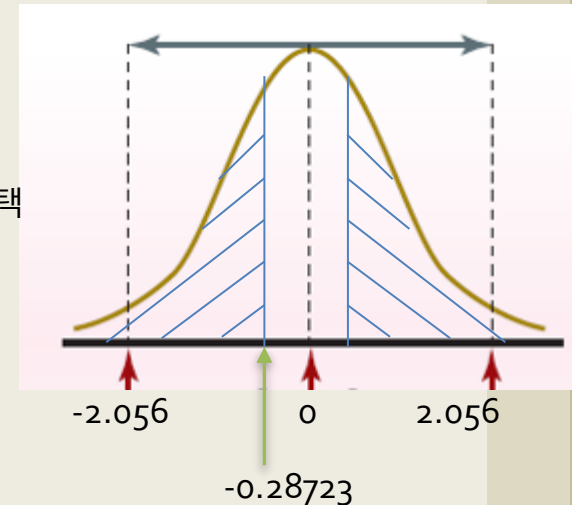


귀무가설 H_0 : 거식증 치료방법으로 Cont 그룹의 전후 몸무게에 변화가 없다.
대립가설 H_1 : 거식증 치료방법으로 Cont 그룹의 전후 몸무게에 변화가 있다.

5%

- paired - t검정, 양측 검정 선택
- t 분포에서의 임계치 계산
→ -2.056, 2.056 (0.05, 25)

- 표본의 t통계량 -0.28723
- p-Value 값 산출 0.7763



검정통계량 -0.28723 과 임계값 -2.056 비교 (귀무가설 기각 범위에 있는지)
p-Value 0.7763 < 0.05 인지

결론 귀무가설 H_0 채택 :
쌍체비교 검정결과 거식증 치료방법으로 Cont를
시행했을 때(아무것도 안 한 그룹) 전후의 몸무게에
변화가 없다고 판단할 수 있다.



연습문제 7.8 - R script

- ❑ anorexia.csv의 Treat에서 “CBT”를 배제하고 Diff(Postwt-Prewt)가 두 Treat(Cont, FT) 간에 차이가 있는지 조사해보자
 - ❑ Boxplot을 그려보자
 - ❑ var.test() 등분산 검정을 해보자
 - ❑ t.test()로 등분산 검정 결과에 맞는 two-sample t-test를 해보자

#7.8 연습문제

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

```
anorexia2 <- anorexia[anorexia$Treat!="CBT",]  
anorexia2$diff <- anorexia2$Postwt-anorexia2$Prewt  
boxplot(diff~Treat, data=anorexia2, col='green')  
var.test(diff~Treat, anorexia2)  
t.test(diff~Treat, var.equal = T, data=anorexia2)
```

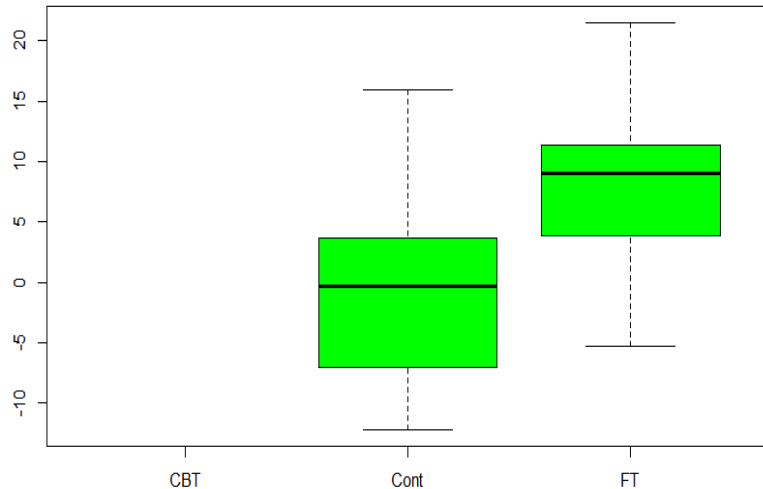
#두 그룹의 분포가 차이가 있는지 시각적으로 확인
#등분산인지 확인 - 분산이 같다

```
# Cont <- anorexia[anorexia$Treat=="Cont",]  
# FT <- anorexia[anorexia$Treat=="FT",]  
# new_anorexia <- rbind(Cont, FT) # cbind는 컬럼 추가  
# new_anorexia$diff <- new_anorexia$Postwt-new_anorexia$Prewt  
# boxplot(diff~Treat, data=new_anorexia, col='green') #두 그룹의 분포가 차이가 있는지 시각적으로 확인  
# var.test(diff~Treat, data=new_anorexia) #등분산인지 확인 - 분산이 같다  
# t.test(diff~Treat, var.equal = T, data=new_anorexia)
```




연습문제 7.8 - R 분석결과

▶ 1. Boxplot



▶ 2. F-test : var.test()

F test to compare two variances

```
data: diff by Treat
F = 1.2458, num df = 25, denom df = 16, p-value = 0.6587
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4766073 2.9699299
sample estimates:
ratio of variances
 1.245775
```

두 모집단의 분산의 비율인 F값이 1.2458
p-value > 0.05 이므로 귀무가설(분산이 같다) 채택

▶ 3. two sample t-test : t.test()

Two Sample t-test

```
data: diff by Treat
t = -3.2227, df = 41, p-value = 0.002491
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.549248 -2.880164
sample estimates:
mean in group Cont    mean in group FT
    -0.450000         7.264706
```

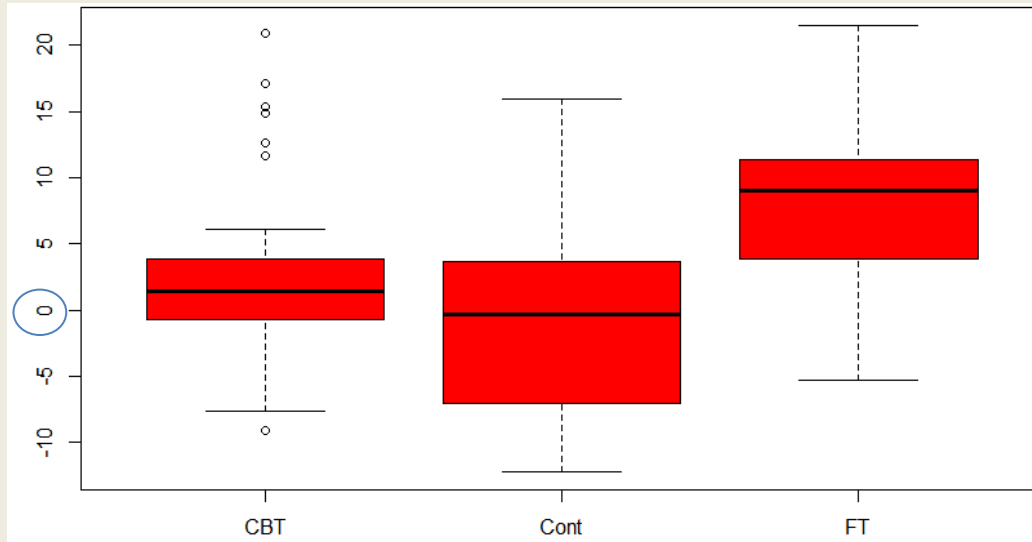
Cont (약을 안 먹은 그룹)의 몸무게 변화의 평균은 -0.45
FT(치료제 FT를 먹은 그룹)의 몸무게 변화의 평균은 7.26

두 집단의 평균의 차이 -7.71의 t 통계량은 -3.2227(자유도 41)
p-value < 0.05 이므로 귀무가설(평균의 차이가 0이다) 기각

치료제 FT의 체중변화가 통계적으로 유의미하다.



anorexia data (전후 몸무게 차이)



- ▶ 각 치료 그룹에서 몸무게 전후 비교 → One sample t-test, paired t-test
 - ▶ Cont : 정규성 만족, t-test결과 차이가 없다는 귀무가설 채택
 - ▶ FT : 정규성 만족, t-test결과 차이가 없다는 귀무가설 기각
 - ▶ CBT : 정규성 만족하지 않음, **Wilcoxon Signed-Rank test** 결과 차이가 없다는 귀무가설 채택
- ▶ 치료 그룹간 몸무게 전후변화 비교 → two sample t-test
 - ▶ Cont vs FT : 등분산, t-test결과 차이가 없다는 귀무가설 기각, **잔차 정규성 만족**
 - ▶ Cont vs CBT : 등분산, t-test결과 차이가 없다는 귀무가설 채택, **잔차 정규성 만족하지 않음**
 - ▶ Two-sample t-test 대신 **Wilcoxon Rank-Sum test** 결과 차이가 있다는 귀무가설 채택
 - ▶ FT vs CBT : 등분산, t-test결과 차이가 없다는 귀무가설 채택, **잔차 정규성 만족**
- ▶ 세 치료 그룹간 몸무게 전후변화 비교 ?? → **분산분석!!!**



anorexia data (전후 몸무게 차이) 분석 R-script

□ Cont vs CBT

```
anorexia3 <- anorexia[anorexia$Treat!="FT",]  
anorexia3$diff <- anorexia3$Postwt-anorexia3$Prewt
```

```
var.test(diff~Treat, anorexia3)  
t.test(diff~Treat, var.equal = T, data=anorexia3)
```

```
out <- lm(diff~Treat, data = anorexia3)  
shapiro.test(resid(out))           # 잔차의 정규성 검증
```

```
wilcox.test(diff~Treat, data = anorexia3)  # 정규성을 만족하지 않음..
```

□ FT vs CBT

```
anorexia4 <- anorexia[anorexia$Treat!="Cont",]  
anorexia4$diff <- anorexia4$Postwt-anorexia4$Prewt
```

```
var.test(diff~Treat, anorexia4)  
t.test(diff~Treat, var.equal = T, data=anorexia4)
```

```
out <- lm(diff~Treat, data = anorexia4)    # 잔차의 정규성 검증  
shapiro.test(resid(out))
```



연습문제 - 상관성 분석

- ❑ SC_survey2021.csv 를 이용하여 다음을 통계적으로 분석하여 정리하여라.
 - ❑ ‘중고등학교 때 통계수업을 좋아하는 정도와 중고등학교 통계성적의 상관성이 있을 것이다’라는 가설의 검정
 - ❑ 임의로 선정한 세 개 그룹에 연령대가 골고루 뿔었는지 동질성 검정

변수명	변수 설명	데이터
Exp_ST	이전에 통계수업을 들었는지	Y, N
Exp_SP	이전에 통계패키지를 사용해본 경험이 있는지	Y, N
Exp_CP	이전에 컴퓨터 프로그램 언어를 배운 경험이 있는지	Y, N
Pref_MH_ST	중고등학교 때 통계수업을 좋아한 정도	1~5
Gr_MH_ST	중고등학교 때 통계수업 성적 정도	1~5
Pref_SP	통계수업이 통계패키지 중심으로 진행되길 원하는 정도	1~5
Grade	이 과목 예상 학점	1~5
AgeGroup	연령그룹	2,3,4,5
Gender	성별	M, F
Group	세 개의 그룹으로 랜덤하게 편성한 그룹	1,2,3



연습문제 - 상관성 분석 결과

#연습문제

```
SC_survey2021 <- read.csv('SC_survey2021.csv')

#중고등학교 때 통계수업을 좋아하는 정도와 중고등학교 통계성적의 상관성
with(SC_survey2021, cor.test(Pref_MH_ST, Gr_MH_ST))

#그룹에 연령그룹이 골고루 있는지
M1=xtabs(~Group + AgeGroup, data = SC_survey2021)
M1
ht.out1 <- chisq.test(M1)
ht.out1
```

□ 상관관계(피어슨)분석 결과

- 상관계수는 0.72로 양의 상관성이 강하게 존재
- p-value가 0.05보다 작으므로 두 변수간의 상관 계수가 0이라는 귀무가설을 기각

```
> with(SC_survey2021, cor.test(Pref_MH_ST, Gr_MH_ST))

Pearson's product-moment correlation

data: Pref_MH_ST and Gr_MH_ST
t = 7.4818, df = 51, p-value = 9.454e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5631903 0.8311532
sample estimates:
      cor
0.7233681
```

□ 교차분석 결과

- 교차분석 결과 p-value가 0.05보다 큰 0.08139 이므로 그룹별 배분된 연령그룹의 분포가 동질성을 가지고 있다고 주장하는 귀무가설을 채택한다.

```
> #그룹에 연령그룹이 골고루 있는지
> M1=xtabs(~Group + AgeGroup, data = SC_survey2021)
> M1
      AgeGroup
Group  2   3   4   5
  1   7   4   5   1
  2   5   8   4   1
  3  13   5   0   0
> ht.out1 <- chisq.test(M1)
경고메시지(들):
In chisq.test(M1) : 카이제곱 approximation은 정확하지 않을수도 있습니다
> ht.out1

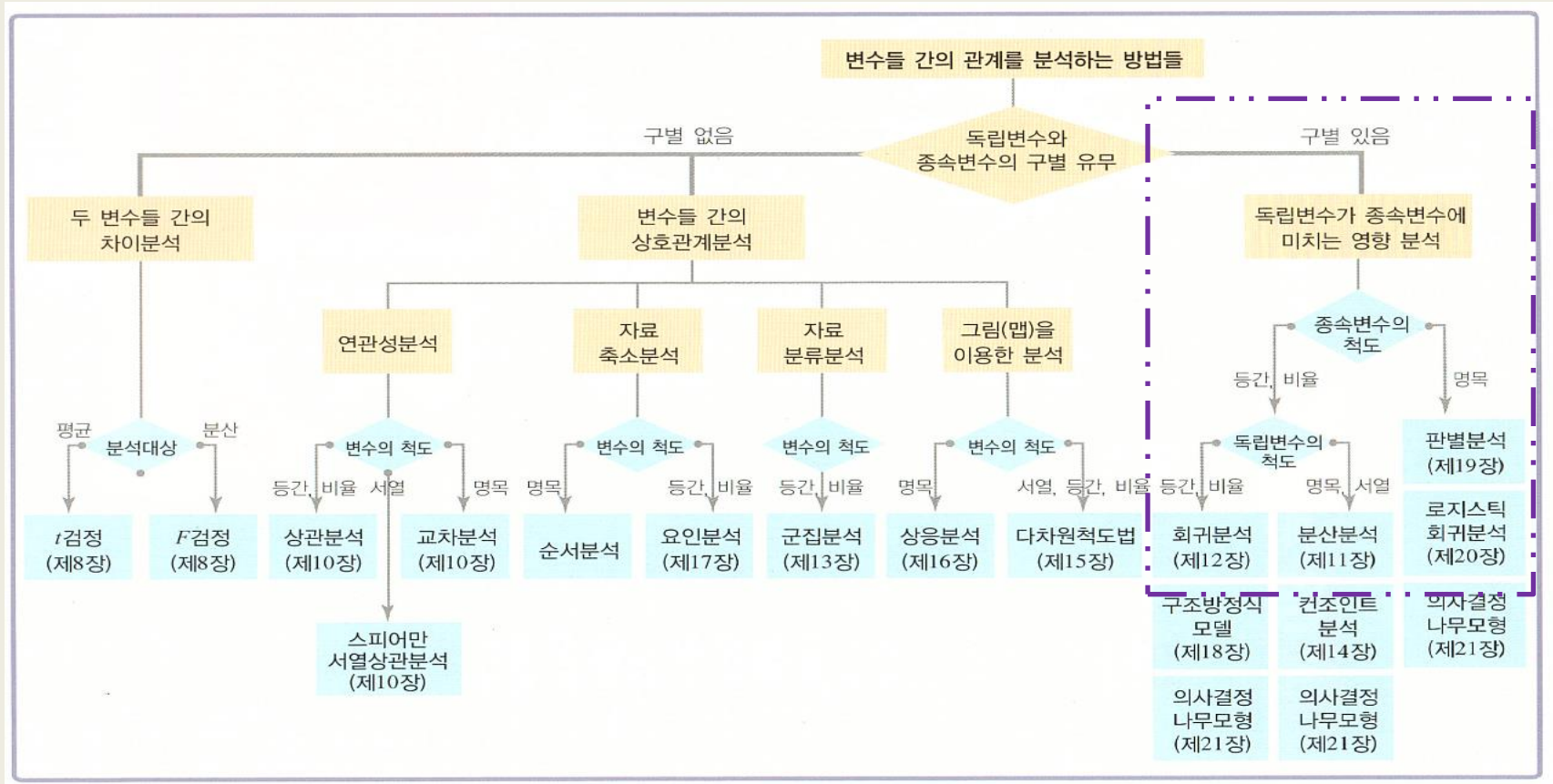
Pearson's Chi-squared test

data: M1
X-squared = 11.235, df = 6, p-value = 0.08139
```

→ 빈도수가 5보다 작은 경우 fisher 정확검정
fisher.test(M1) 결과 p-value가 0.05보다 작은 0.03599

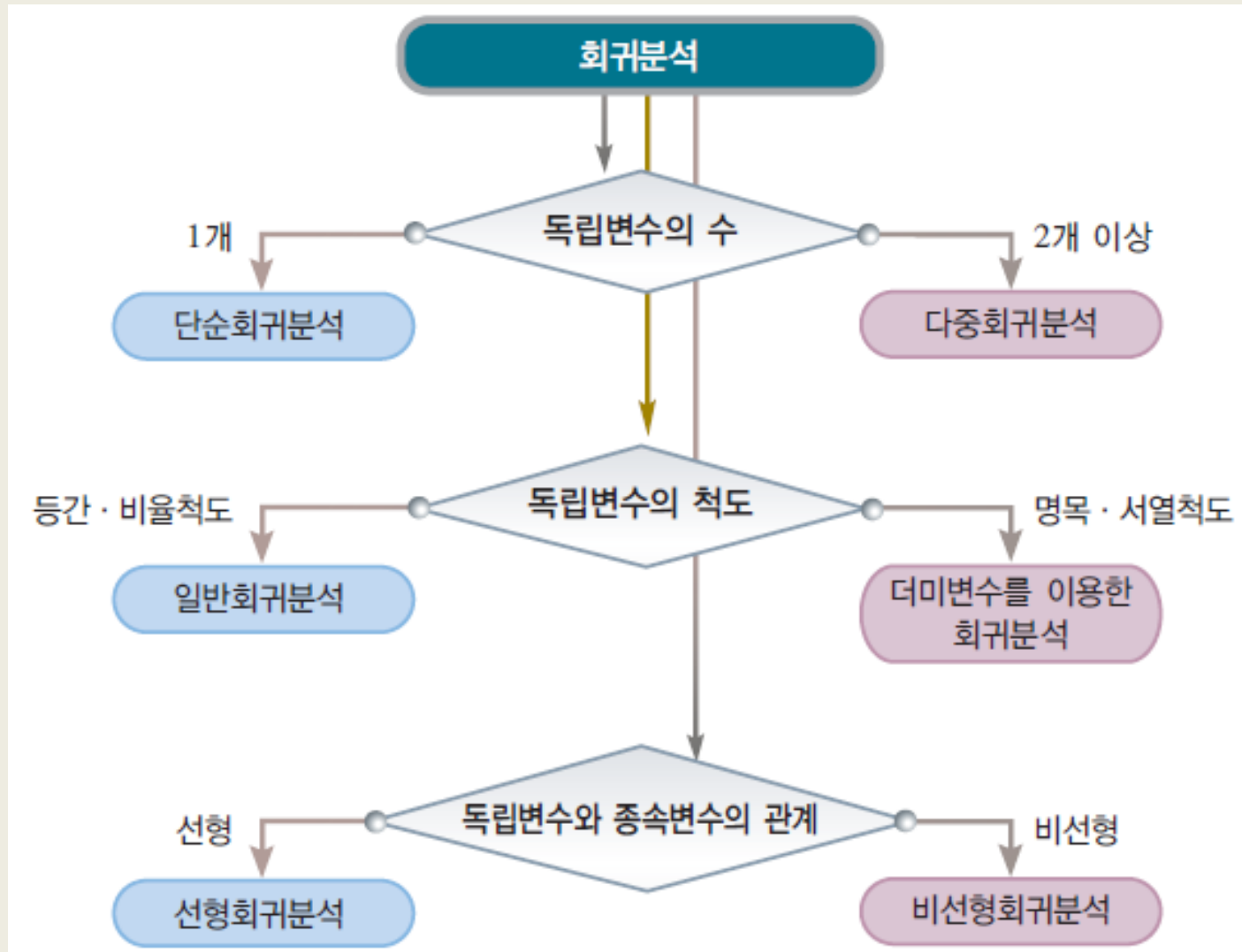


분석방법의 선택





회귀분석의 종류





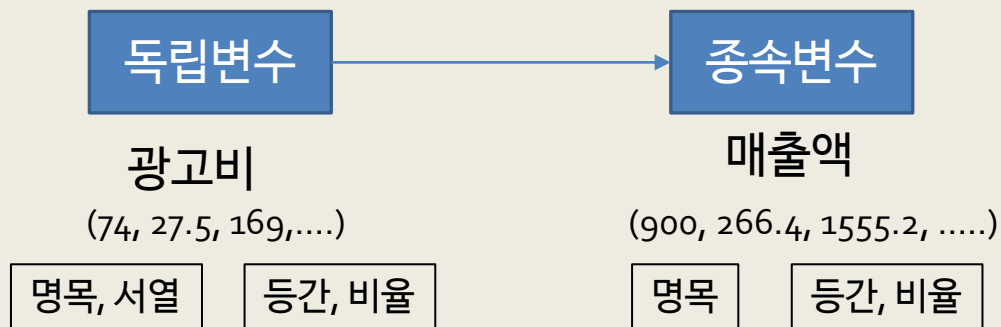
독립변수와 종속변수의 관계분석 사례 1

-적절한 분석방법은?

- ❑ 어느 회사의 담당자는 내년 마케팅 계획을 준비하기 위해서 제품수요 예측을 하려고 한다.
- ❑ 제품수요를 가장 잘 예측할 수 있는 것이 광고비 지출이라고 보고 8개 테스트 시장에서 매출액과 광고비 지출 내역을 수집하였다.

- ❑ 자료 : 8개 테스트 시장에서 수집된 자료

매출액	광고비
900	74
266.4	27.5
1555.2	169
4320	497
2707.2	270.5
439.2	44.5
1209.6	63
2966.4	189.5



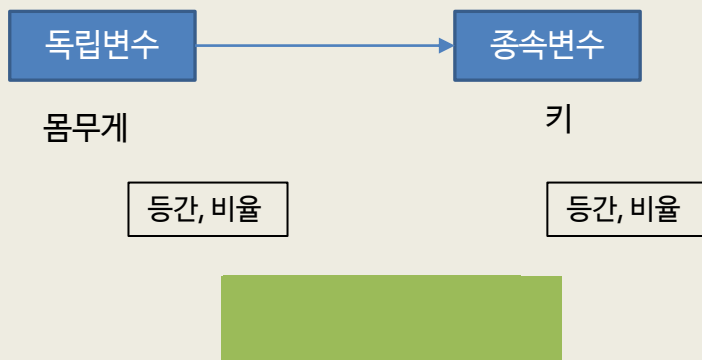


독립변수와 종속변수의 관계분석 사례 2

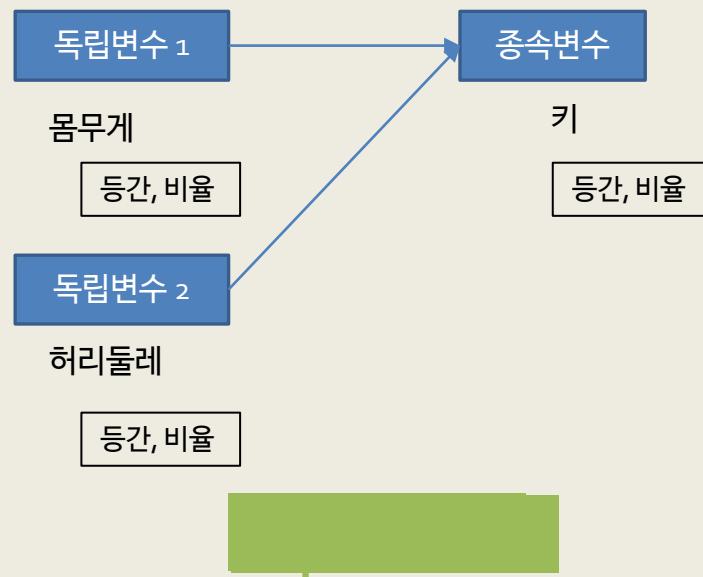
-적절한 분석방법은?

- 어느 고등학교 A반에 한 학생이 전학 오기로 했다. 이 학생의 키가 얼마나 되는지 추정하려고 한다. 몸무게가 70kg이라고 한다. 다음의 두 가지 방법으로 추정하였다.
 - 몸무게만으로 키를 추정하는 방법(방법 1)
 - 몸무게와 허리 둘레로 키를 추정하는 방법(방법 2)
- 반 학생 50명의 키, 몸무게, 허리둘레를 조사한 자료

□ 방법 1



□ 방법 2





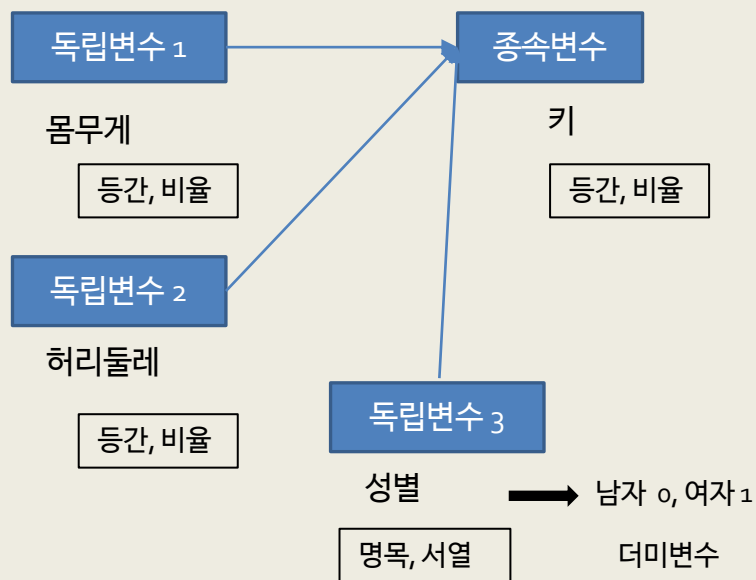
독립변수와 종속변수의 관계분석 사례 3

-적절한 분석방법은?

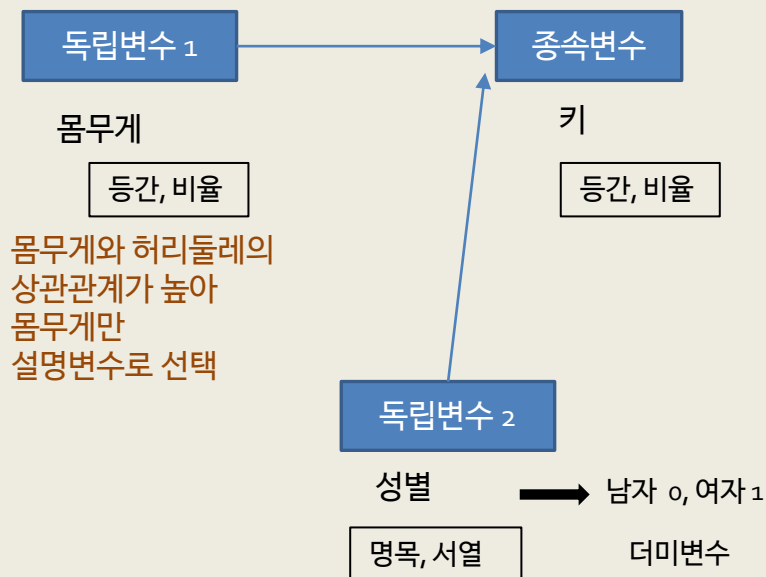
- 어느 고등학교 A반에 한 학생이 전학오기로 했다. 이 학생의 키가 얼마나 되는지 추정하려고 한다. 몸무게가 70kg이라고 한다. 다음의 두 가지 방법으로 추정하였다.
 - 몸무게와 허리 둘레로 키를 추정하는 과정에서 남녀 간의 차이가 분명하게 있음을 발견하여 성별을 키를 추정하는 요소에 포함시키고자 하였다.

- 반 학생 50명의 키, 몸무게, 허리둘레, 성별을 조사한 자료

□ 방법 3



□ 방법 4

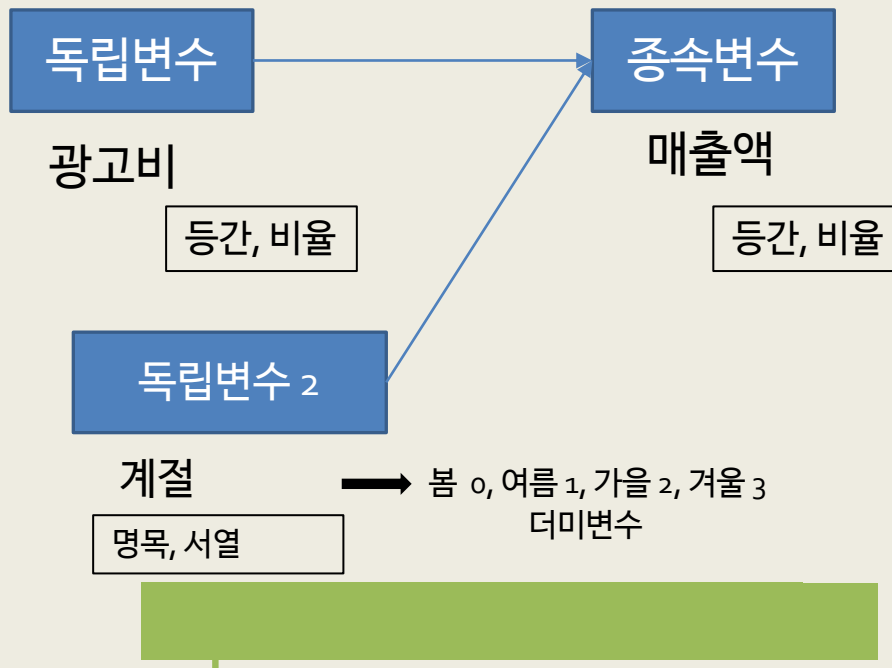




독립변수와 종속변수의 관계분석 사례 4

-적절한 분석방법은?

- ❑ 어느 회사의 담당자는 내년 마케팅 계획을 준비하기 위해서 수요예측을 하려고 한다.
 - ❑ 제품수요를 가장 잘 예측할 수 있는 것이 광고비 지출이라고 보고 8개 테스트 시장에서 매출액과 광고비 지출 내역을 수집하였다. (사례 1 참조)
 - ❑ 매출액은 계절과도 영향이 있다고 판단하여 계절도 예측요소에 포함하였다.
- ❑ 자료 : 8개 테스트 시장에서 수집된 자료를 계절별로 세분함





독립변수와 종속변수의 관계분석 사례 4 -더미변수를 이용한 회귀분석 예

더미변수를 이용한 회귀분석

S전자는 계절과 광고비에 따라 매출액의 차이가 어느 정도 나는지 알아보기 위해서 회귀분석을 실시하고자 하였다. 그러나 계절은 명목척도로 측정되어 독립변수로 사용할 수 없다. 따라서 계절변수를 3개의 이항변수(더미변수)로 더미코딩하여 회귀분석을 실시하고자 다음과 같이 준비하였다.

더미변수를 이용해 추정한 회귀식

〈잘못된 일반적인 회귀식〉

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

〈더미변수를 이용한 회귀식〉

$$\hat{y}_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 x_{4i}$$

여기서, \hat{y}_i : 추정된 매출액(단위: 억원)

x_{1i} : 계절을 나타내는 변수

x_{2i}, x_{4i} : 광고비(단위: 억원)

D_{1i}, D_{2i}, D_{3i} : 계절을 나타내는 더미변수

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$: 추정되어야 할 모수들

■ 4개의 수준을 갖는 계절변수를 봄을 기준으로 하여 3개의 더미변수로 코딩하는 방법

계절	더미변수 1(D_{1i})	더미변수 2(D_{2i})	더미변수 3(D_{3i})
봄	0	0	0
여름	1	0	0
가을	0	1	0
겨울	0	0	1



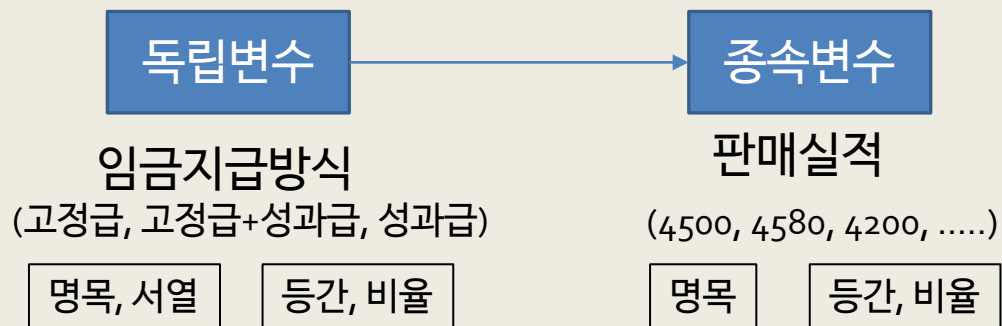
독립변수와 종속변수의 관계분석 사례 5

-적절한 분석방법은?

- ❑ 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.

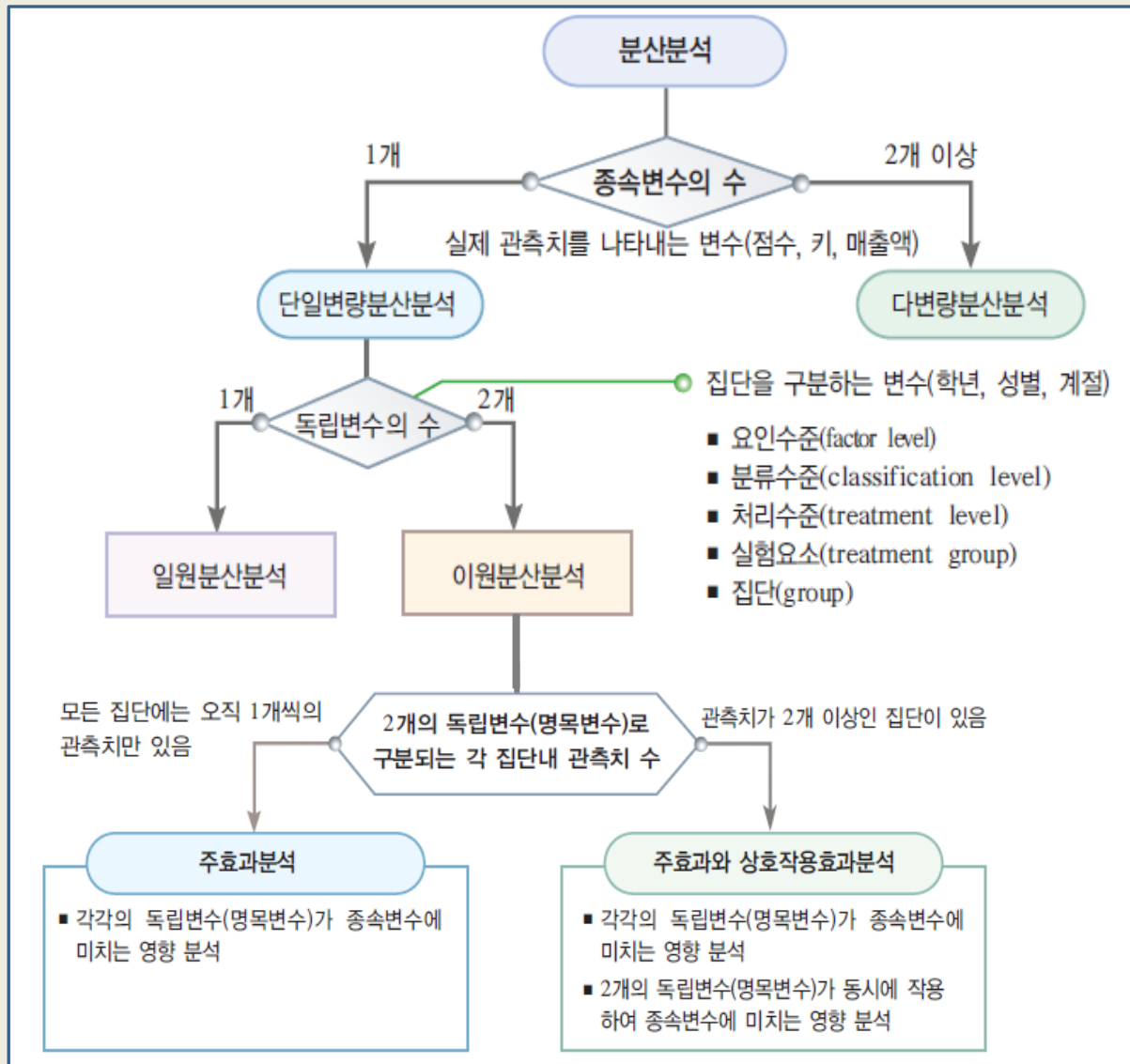
■ 자료 : 6개월간 24명의 판매실적(각 대안별 8명)

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





분산분석의 종류





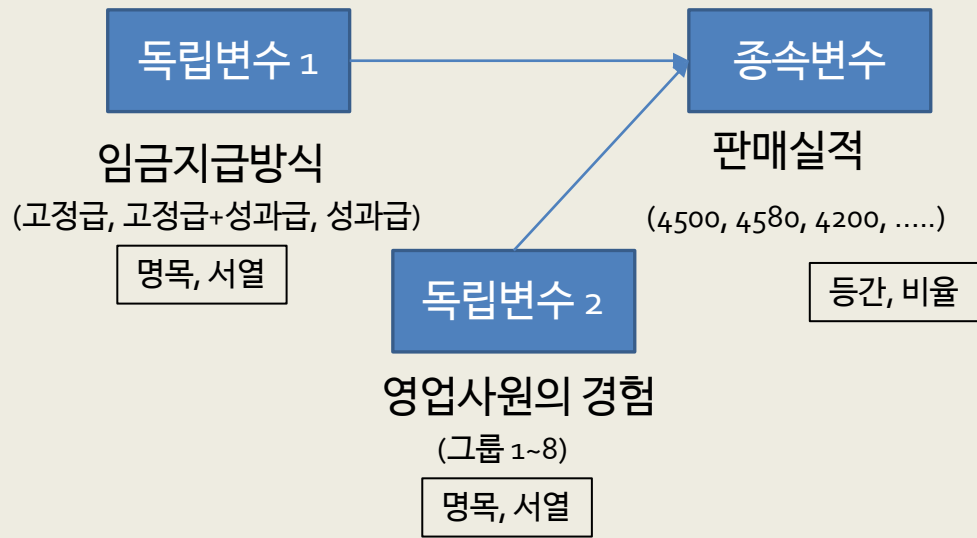
독립변수와 종속변수의 관계분석 사례 6

-적절한 분석방법은?

- ❑ 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	FixPlncen	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



*모든 집단에 오직 1개씩의 관측치만 있어
각각의 독립변수가 종속변수에 미치는 영향을 분석
→ 주효과 분석 또는 반복 없는 이원 배치법



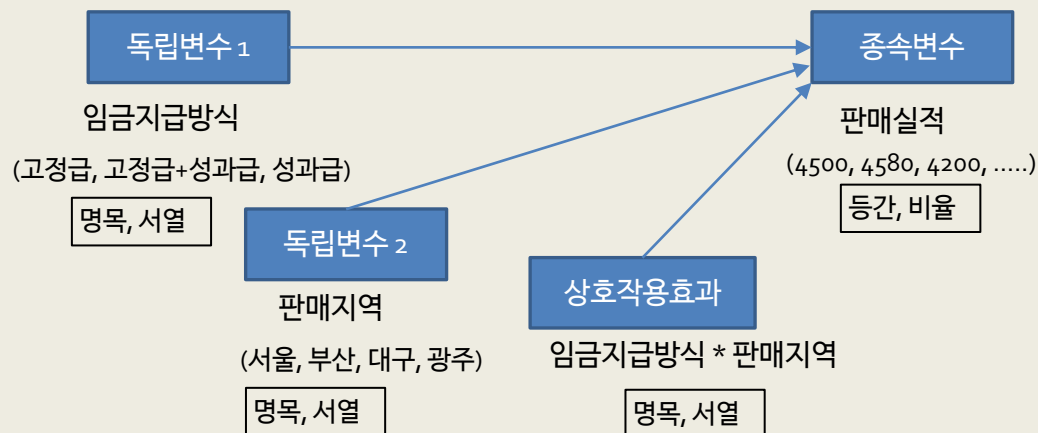
독립변수와 종속변수의 관계분석 사례 7

-적절한 분석방법은?

- 어느 문구판매전문회사에서 임금지급방식에 대해 고민하고 있다. (1) 고정급만 받는 방식, (2) 고정급과 성과급을 함께 받는 방식, (3) 성과급만 받는 방식 세 가지이다.
- 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



*관측치가 2개 이상인 집단이 있어
 각각의 독립변수가 종속변수에 미치는 영향을 분석하고
 2개의 독립변수가 동시에 작용하여 종속변수에 미치는 영향 분석
 → 주효과와 상호작용효과 분석(반복 있는 이원 배치법)



독립변수와 종속변수의 관계분석 사례 8

-적절한 분석방법은?

□ 다음은 대통령 후보자 선택 시 후보의 성실성과 개혁성향에 대해 어느 정도 중요하게 생각하는지에 대한 설문조사 양식과 응답자료이다.

■ 대통령 후보자를 선택할 때 당신은 다음 요인들을 어느 정도로 중요하게 생각하십니까?

1) 후보자의 성실성

① ② ③ ④ ⑤
전혀 중요하지 보통 매우 중요함
않음

2) 후보자의 개혁성향

① ② ③ ④ ⑤
전혀 중요하지 보통 매우 중요함
않음

■ 다음의 후보자 중에서 어느 후보를 더 선호하십니까?

① 후보 1



② 후보 2



유권자	성실성에 대한 중요도	개혁성향에 대한 중요도	선호후보자
1	4	3	2
2	5	1	2
3	4	1	2
4	1	2	1
5	1	2	2
6	2	1	2
7	3	4	1
8	3	2	2
9	2	1	2
10	2	1	1
⋮	⋮	⋮	⋮

독립변수 1

성실성에 대한 중요도
(1~5)

등간, 비율

독립변수 2

개혁성향에 대한 중요도
(1~5)

등간, 비율

종속변수

선호후보자
(후보 1, 후보 2)

명목



*독립변수들이
연속변수여야 함



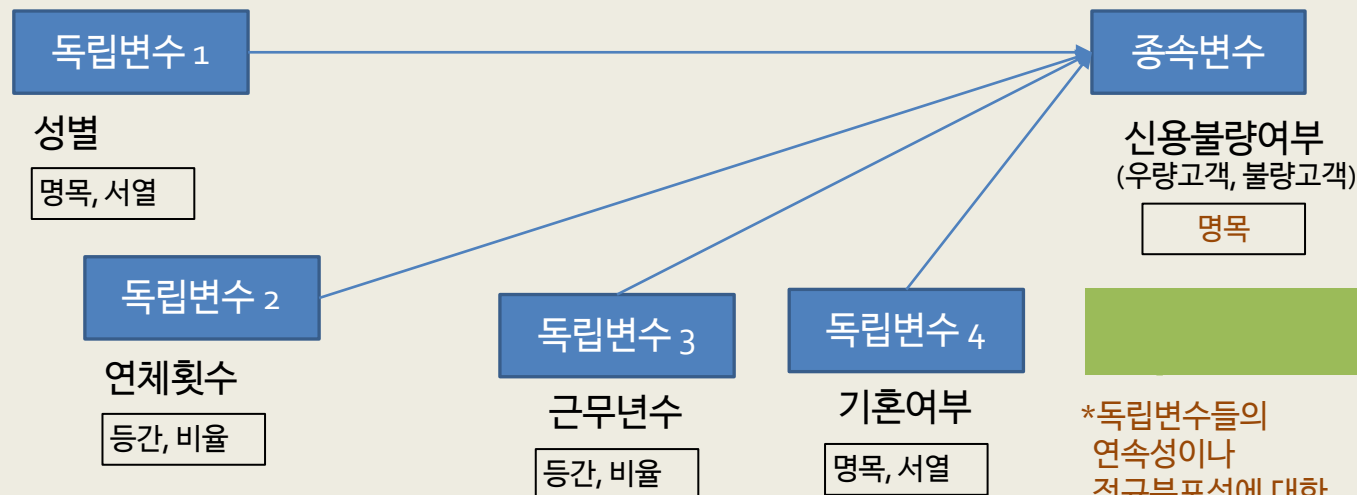
독립변수와 종속변수의 관계분석 사례 9

-적절한 분석방법은?

■ K은행에서 고객의 성별, 연체횟수, 직장근무년수, 기혼여부 등에 따라 고객의 신용상태를 추정할 수 있는 로지스틱 회귀모형을 개발하고자 다음과 같은 자료를 수집함

- 1) 고객의 성별? ①  ② 
- 2) 연체횟수? ()회
- 3) 직장근무년수? ()년
- 4) 기혼여부? ① 미혼 ② 기혼
- 5) 고객의 신용상태?
① 우량고객 ② 불량고객

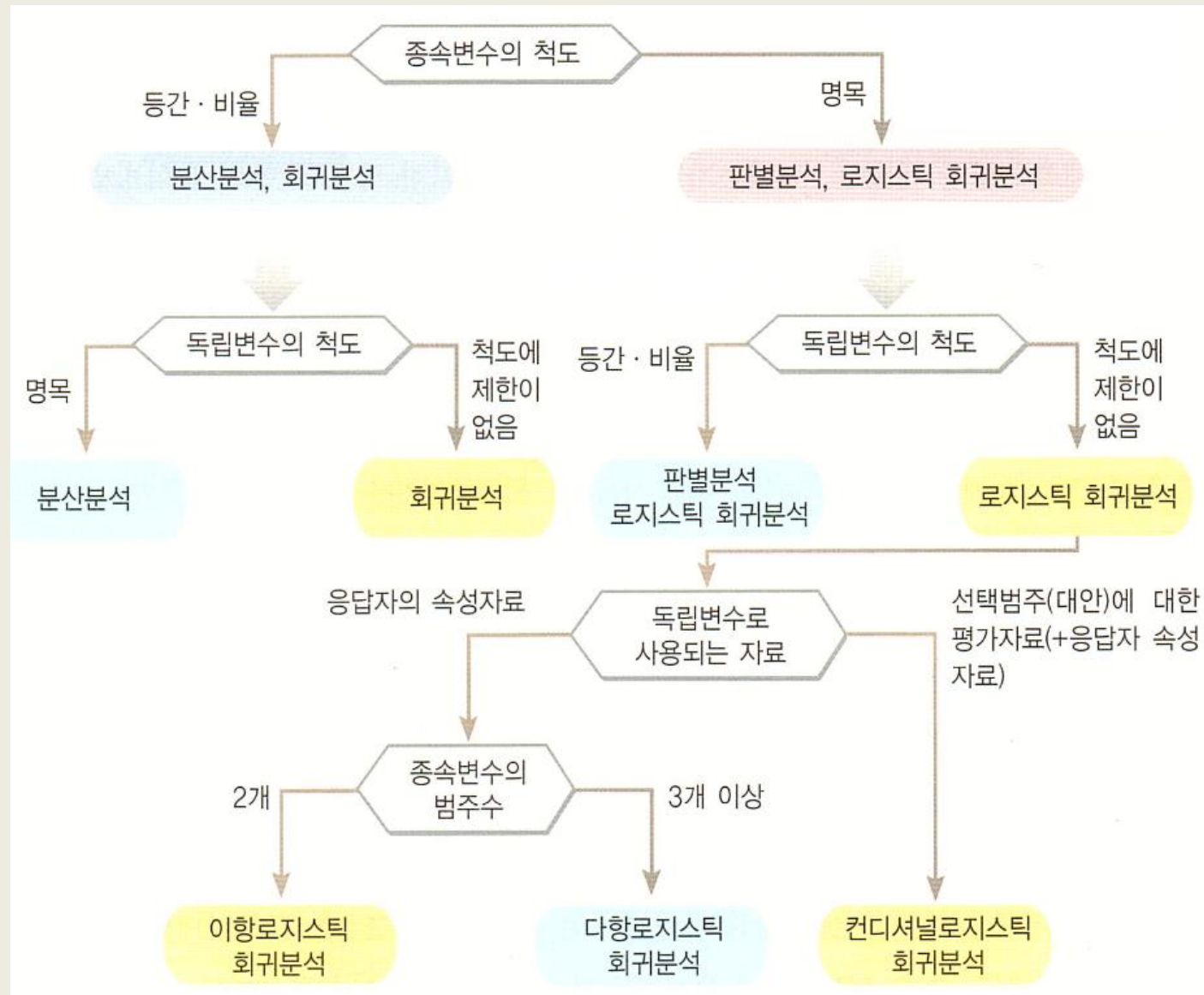
성별	연체횟수	근무년수	기혼여부	신용불량여부
1	12	1	1	1
2	15	3	2	0
2	6	1	2	1
1	35	3	1	0
1	60	4	1	1
2	14	2	2	0
2	38	5	1	1
2	20	5	2	0
1	38	3	1	0
1	57	1	1	1
⋮	⋮	⋮	⋮	⋮



*독립변수들의 연속성이나 정규분포성에 대한 제한이 없음



종속변수를 예측하는 분석방법의 선택

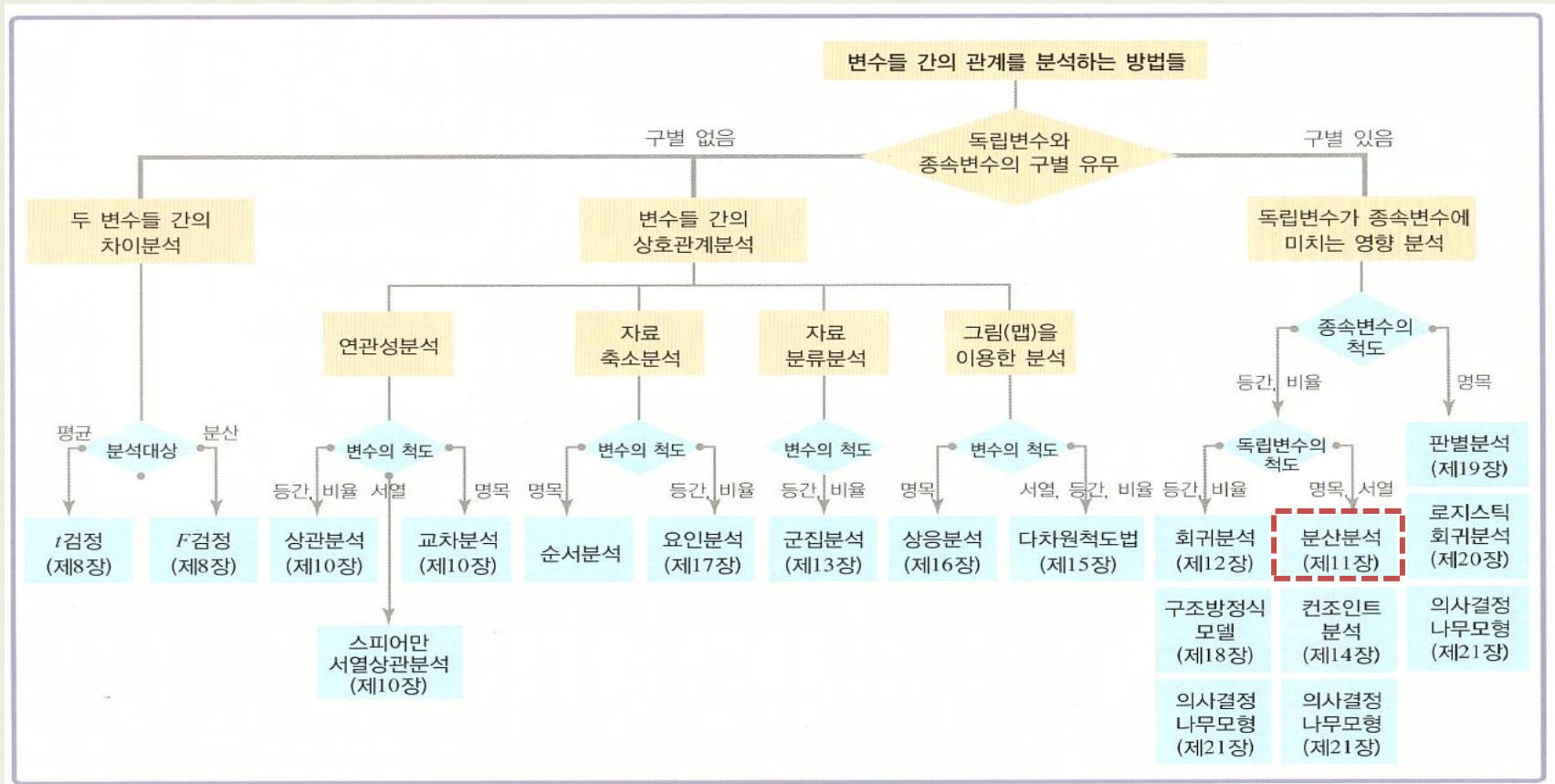




분산분석

(ANOVA, ANalysis Of VAriance)

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





집단간 평균의 분산으로 집단간 평균을 비교

분산분석

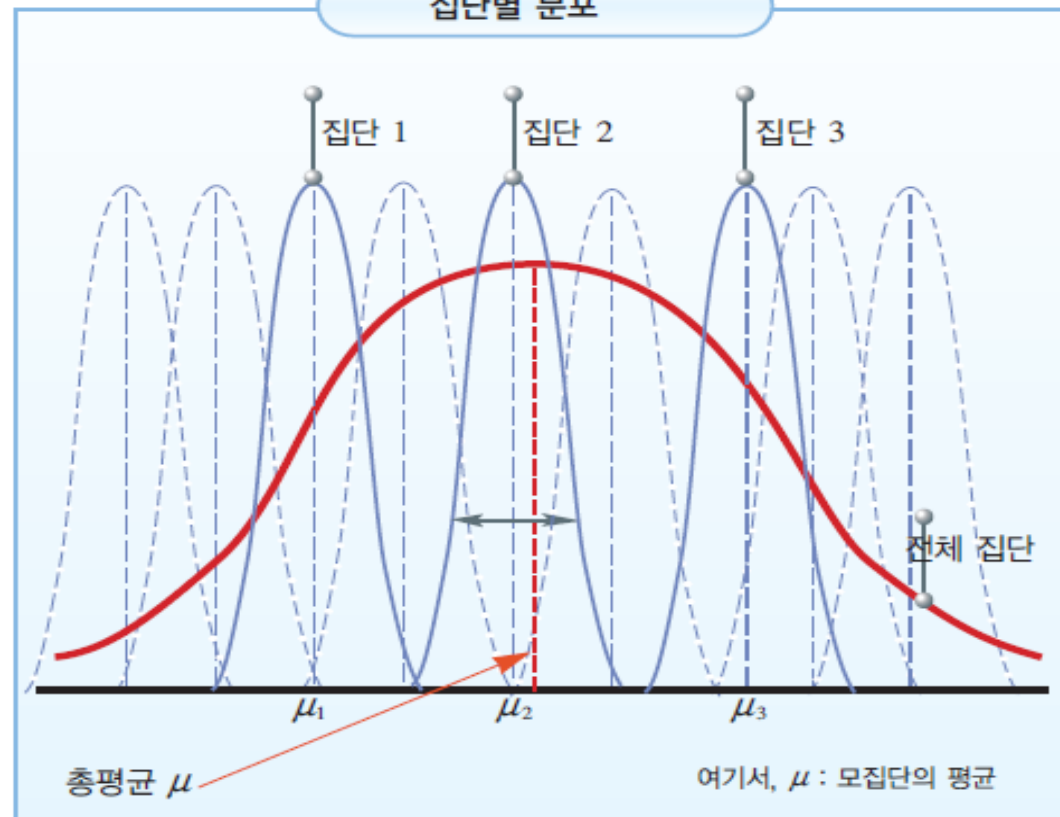
- 집단 간의 평균 차이를 비교하는데 왜 분산을 이용하여 분석하고, 이를 분산분석이라 하는가?

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

전체 집단



집단별 분포



<그림11-2>



평균차이를 분산을 이용하여 분석

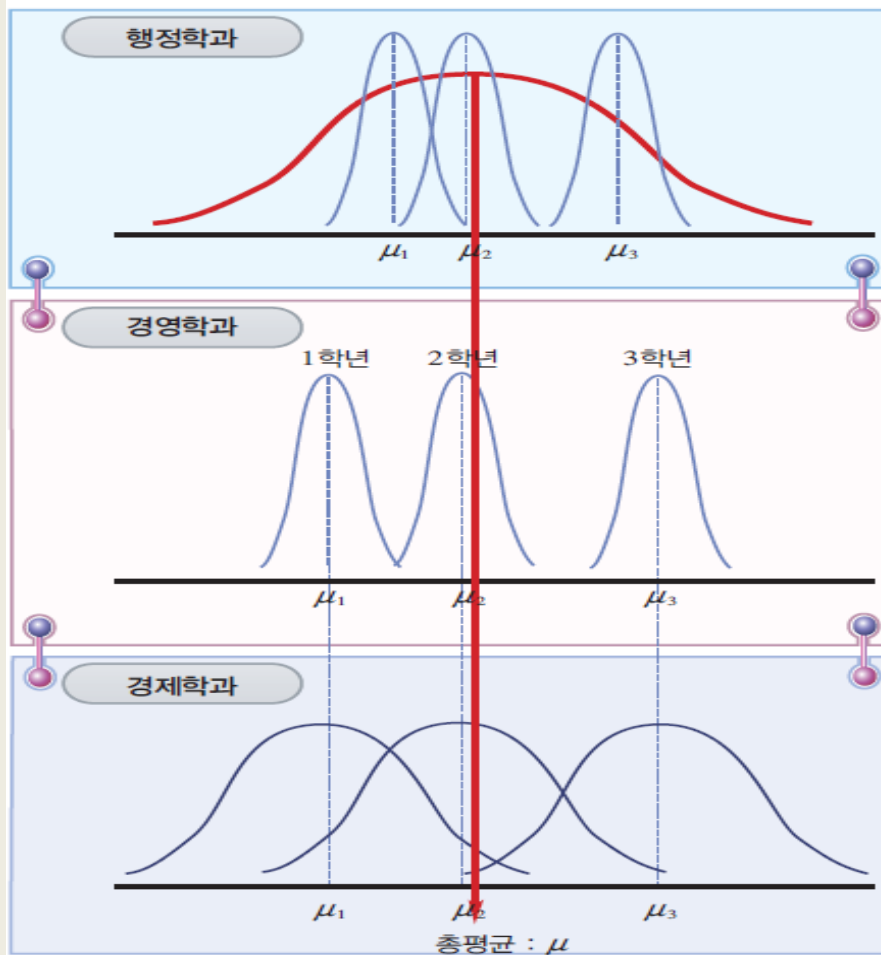
왜 분산분석인가?

● 집단 간의 평균을 비교하는 데 어떻게 분산을 이용하여 분석할 수 있는가?

- 행정학과와 경영학과의 학년별 점수의 분산은 서로 같음
- 경영학과와 경제학과의 학년별 평균 점수는 서로 같음
- 행정학과는 학년별로 평균 점수가 서로 다른가?
- 경영학과는 학년별로 평균 점수가 서로 다른가?
- 경제학과는 학년별로 평균 점수가 서로 다른가?

집단 간 평균 비교의 요소

- 집단 간 평균의 분산
- 집단 내 분산



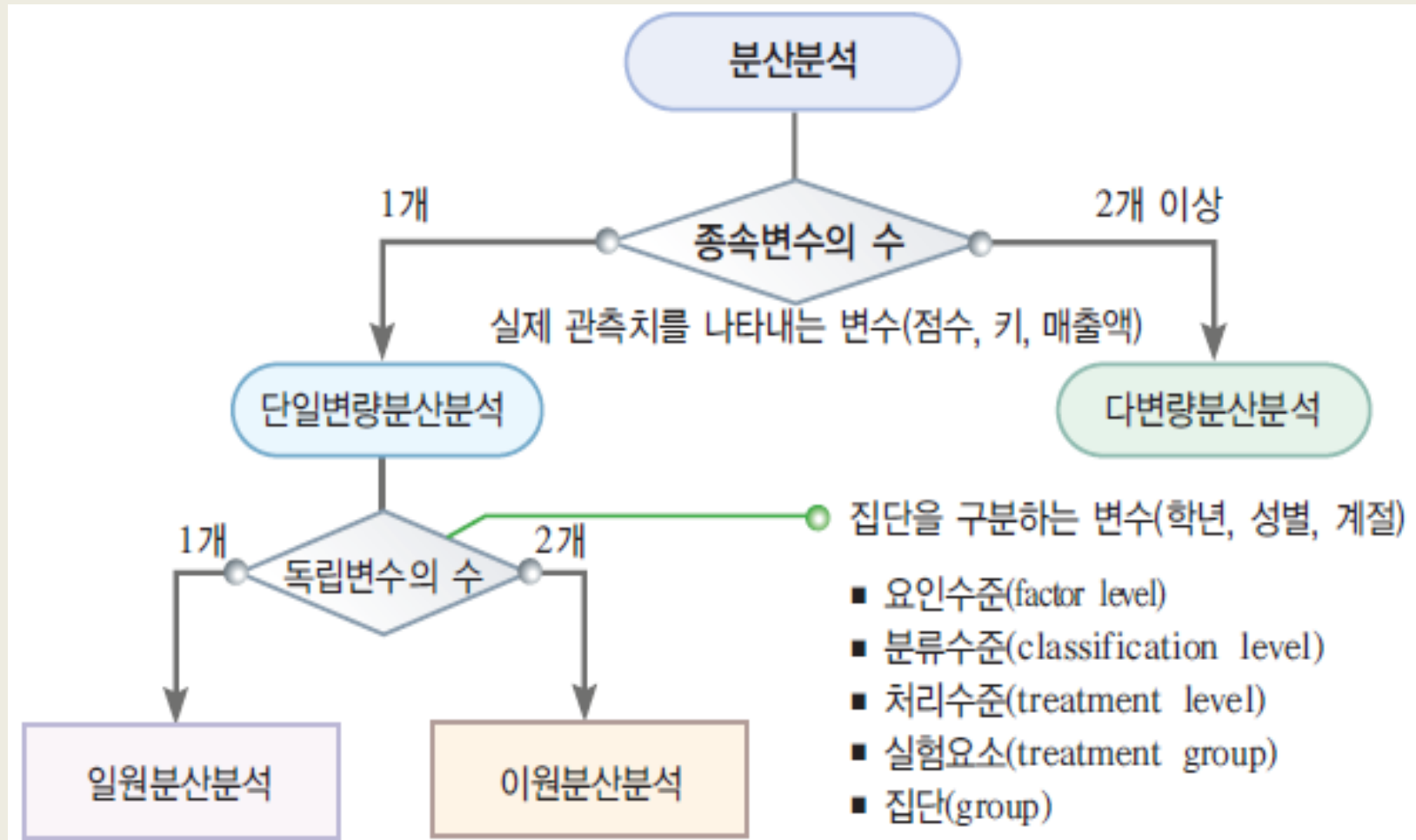
평균차이 결과

- 경영학과는 학년별로 평균 점수에 차이가 있음
- 경제학과는 학년별로 평균 점수에 차이가 있다고 단정하지 못함
 - 학년별 분포에서 많은 부분이 서로 중복됨
 - 3학년 중에는 2학년보다 점수가 나쁜 학생들이 적지 않고, 2학년 중에도 1학년보다 점수가 나쁜 학생들이 많이 있음
- 평균들 간의 차이, 즉 분산뿐만 아니라 집단에 속한 값들의 집단 내 분산이 집단 간 평균의 차이에 대한 판단에 영향을 미치고 있음
- 집단 평균들 간의 분산이 크면 클수록 반면에 집단 내 분산은 작으면 작을수록, 집단 간 평균의 차이가 분명함을 알 수 있음
- 행정학과는 경영학과와 같이 집단 내 분산이 작으나 집단 평균들 간의 분산 또한 상당히 작아서 경제학과와 마찬가지로 학년별 분포에서 많은 부분이 서로 중복됨
- 따라서 행정학과는 학년별로 평균 점수가 명확하게 다르다고 단정할 수 없음

〈그림11-3〉



분산분석의 종류



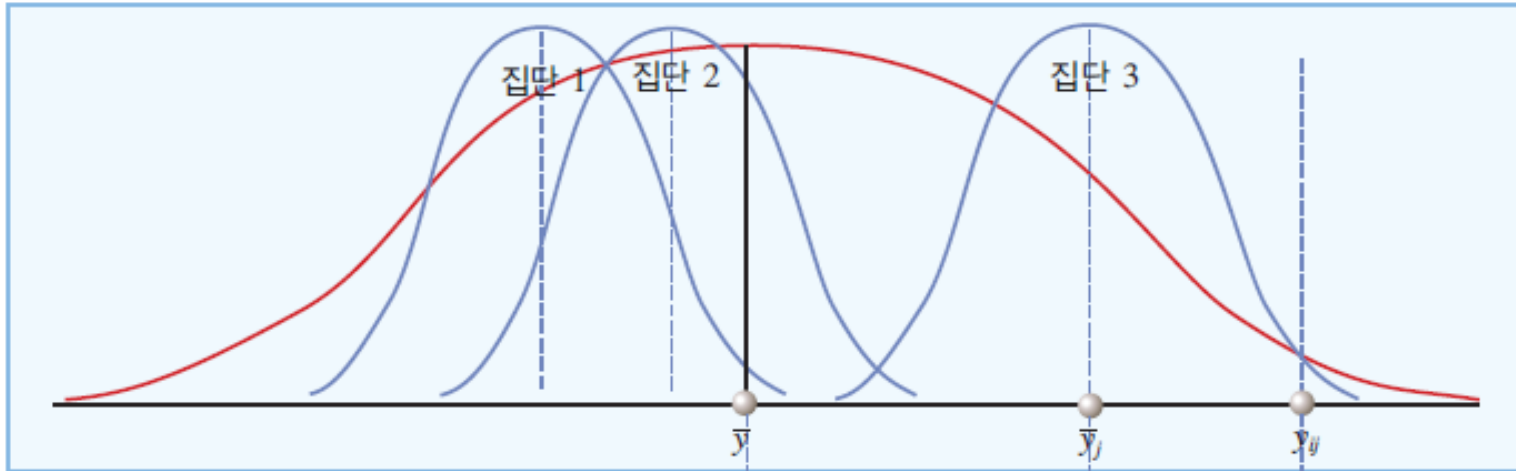
〈그림11-4〉



일원분산분석(One Way ANOVA)

일원분산분석의 개념

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

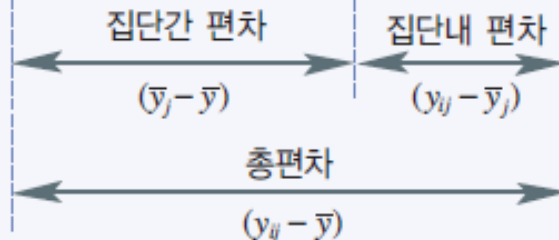


$$y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j)$$

관측치 = 전체평균 + 집단 j 의 처리효과 + 오차

i : 특정한 집단에 속한 관측치를 지정하는 첨자

j : 특정한 집단을 지정하는 첨자



<그림11-5>



일원분산분석

- 편차, 제곱합, 자유도, 평균제곱

한 집단

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

총편차 = 집단간 편차 + 집단내 편차

- 관측치(y_{ij})가 전체평균(\bar{y})으로부터 떨어져 있는 거리($y_{ij} - \bar{y}$)는 전체평균(\bar{y})으로부터 이 관측치(y_{ij})가 속해 있는 집단의 평균(\bar{y}_j)까지의 거리($\bar{y}_j - \bar{y}$)와 관측치(y_{ij})가 속한 집단의 평균(\bar{y}_j)에서부터 관측치(y_{ij})까지의 거리($y_{ij} - \bar{y}_j$)로 구성됨

여러 개 집단

$$\sum_j \sum_i n_{ij} (y_{ij} - \bar{y})^2 = \sum_j \sum_i n_{ij} (\bar{y}_j - \bar{y})^2 + \sum_j \sum_i n_{ij} (y_{ij} - \bar{y}_j)^2$$

총제곱합(SST) = 집단간 제곱합(SSB) + 집단내 제곱합(SSW)

- 분산분석은 식의 우측에 있는 2개 항의 제곱합 크기를 비교하여 집단간 평균이 통계적으로 유의하게 차이가 있는지를 분석함

$$\begin{array}{ccc} (n-1) & (g-1) & (n-g) \\ \hline \text{---} & \text{---} & \text{---} \\ \text{(총제곱합의 자유도)} & = & \text{(집단간 제곱합의 자유도)} + \text{(집단내 제곱합의 자유도)} \end{array}$$



$$\frac{\sum_j \sum_i n_{ij} (y_{ij} - \bar{y})^2}{n-1}$$

$$\frac{\sum_j \sum_i n_{ij} (\bar{y}_j - \bar{y})^2}{g-1}$$

$$\frac{\sum_j \sum_i n_{ij} (y_{ij} - \bar{y}_j)^2}{n-g}$$

(총평균제곱 : MST) (집단간 평균제곱 : MSB) (집단내 평균제곱 : MSW)
(총분산) (집단간 분산) (집단내 분산)

여기서, n : 전체 관측치 수
 g : 집단 수

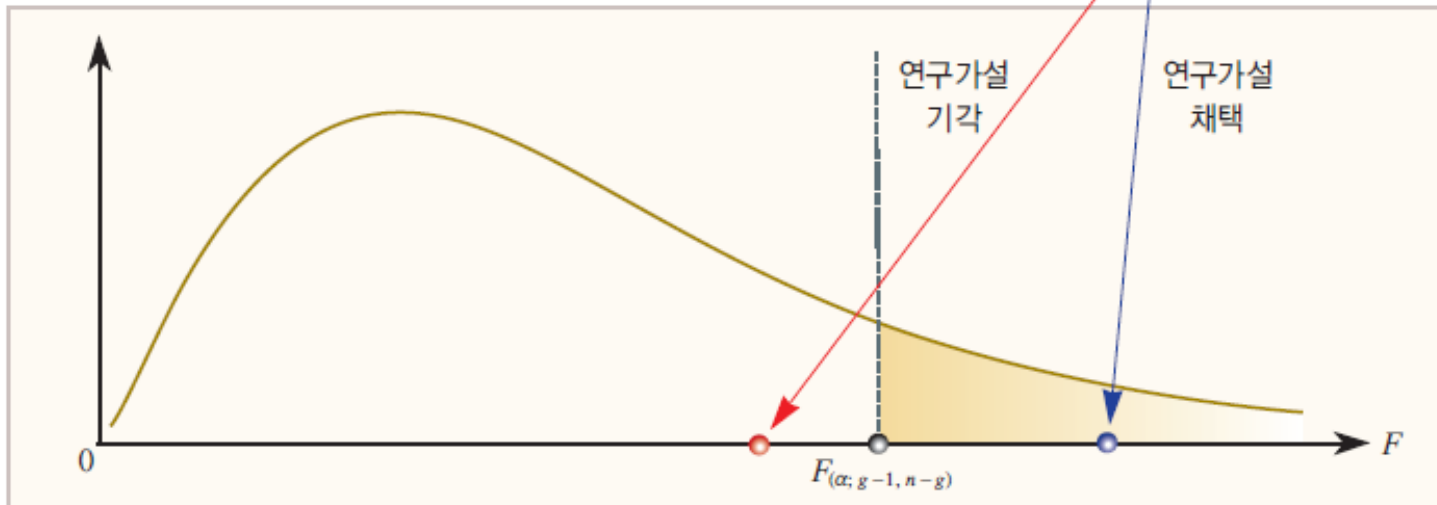
- 평균제곱은 각각의 제곱합을 해당자유도로 나눈 값임
- 제곱합은 관측치의 수(n_{ij})와 집단의 수(g)에 따라 크기가 달라지므로 제곱합을 사용하 기보다는 이들의 수에 영향을 받지 않는 평균제곱을 사용하여 분석해야 함

<그림11-6,7,8>



일원분산분석표와 유의성 검정

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)	$SST = \sum_j \sum_i (y_{ij} - \bar{y})^2$	$(n-1)$		



- 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



일원분산분석 사례-Excel

- 어느 문구판매전문회사에서 임금지급방식에 대해 고민하고 있다. (1) 고정급만 받는 방식, (2) 고정급과 성과급을 함께 받는 방식, (3) 성과급만 받는 방식 세 가지이다.
- 임금지급방식에 따라 판매실적이 달라질 것이라고 생각하고 있는데 통계적으로 유의한 지 판단하고자 하였다.
- 자료 : 6개월간 24명의 판매실적(각 대안별 8명)

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

분산분석결과

분산 분석: 일원 배치법							
요약표							
인자의 수준	관측수	합	평균	분산			
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	
처리	2296233	2	1148117	6.881303	0.005031	3.4668	
잔차	3503763	21	166845.8				
계	5799996	23					



일원분산분석 사례 - R 활용

■ salesRecord1.xlsx

	A	B	C
1	Fixed	FixPlncn	Incentive
2	4500	4430	5810
3	4580	4740	5420
4	4200	4530	4800
5	4860	4830	5100
6	5040	5100	5460
7	4740	4920	6180
8	4320	4140	4680
9	4410	4320	4620



	A	B
1	payType	SalesRec
2	F	4500
3	F	4580
4	F	4200
5	F	4860
6	F	5040
7	F	4740
8	F	4320
9	F	4410
10	FI	4430
11	FI	4740
12	FI	4530
13	FI	4830
14	FI	5100
15	FI	4920
16	FI	4140
17	FI	4320
18	I	5810
19	I	5420
20	I	4800
21	I	5100
22	I	5460
23	I	6180
24	I	4680
25	I	4620

```

1 #import data file
2
3 library(readxl)
4 salesRecord1 <- read_excel("salesRecord1.xlsx")
5 view(salesRecord1)
6
7 # encode a vector as a factor
8 PT <- factor(salesRecord1$payType)
9 salesRecord1$PT<- factor(salesRecord1$payType)
10
11 # lm(), anova(), aov()
12 gc.out1 <- lm(SalesRec ~ PT, data = salesRecord1)
13 anova(gc.out1)
14
15 aov.out1 <- aov(SalesRec ~ PT, data = salesRecord1)
16 summary(aov.out1)
17

```



```

> gc.out1 <- lm(SalesRec ~ PT, data = salesRecord1)
> anova(gc.out1)
Analysis of Variance Table

Response: SalesRec
          Df Sum Sq Mean Sq F value    Pr(>F)
PT          2 2296233 1148117   6.8813 0.005031 **
Residuals 21 3503762  166846
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> aov.out1 <- aov(SalesRec ~ PT, data = salesRecord1)
> summary(aov.out1)
          Df Sum Sq Mean Sq F value    Pr(>F)
PT          2 2296233 1148117   6.881 0.00503 **
Residuals 21 3503762  166846
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```



분산 분석의 사후검정과 비모수 검정

□ 사후검정

- 분산분석의 결과는 집단들 간의 평균차이가 있다는 것만 알려주고 어떤 집단들 간의 차이가 있는 지의 정보는 제공하지 않음
- 사후검정은 분산분석의 보완분석으로 두 집단 간의 비교를 통하여 집단들 간의 차이 여부를 검정함
 - Duncan 검정 : 사회과학, 관대, 집단의 수가 같을 때
 - Tukey 검정 : 자연과학, 엄격, 집단의 수가 같을 때
 - Scheffe 검정 : 사회과학, 집단의 수가 다를 때

□ Kruskal-Wallis Test : 일원분산분석의 비모수 검정

- `Kruskal.test (y~그룹변수, data = ' ')`



R 실습

- ❑ 누구나 Chapter 12,13
- ❑ 연습문제 12.7
 - ▣ anorexia.csv의 세 그룹 간의 Prewt평균의 차이가 있는지 일원분석을 해보자
 - ▣ anorexia.csv의 세 그룹 간의 전후 몸무게 차이가 있는지 일원분석을 한 후 사후검정을 해보자



PlantGrowth.csv 의 데이터

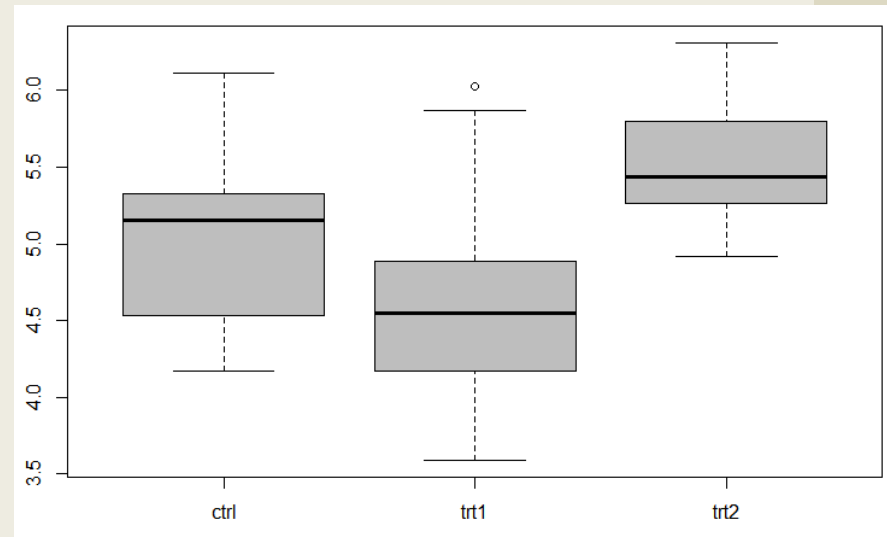
- 종속변수 : weight (수확된 식물의 무게)
- 독립변수 : group (식물이 자라는 조건)
 - Ctrl (대조군), trt1 (실험군 1), trt2 (실험군 2)

- 그룹별 몸무게 평균

ctrl	trt1	trt2
5.032	4.661	5.526

- 그룹별 몸무게 표준편차

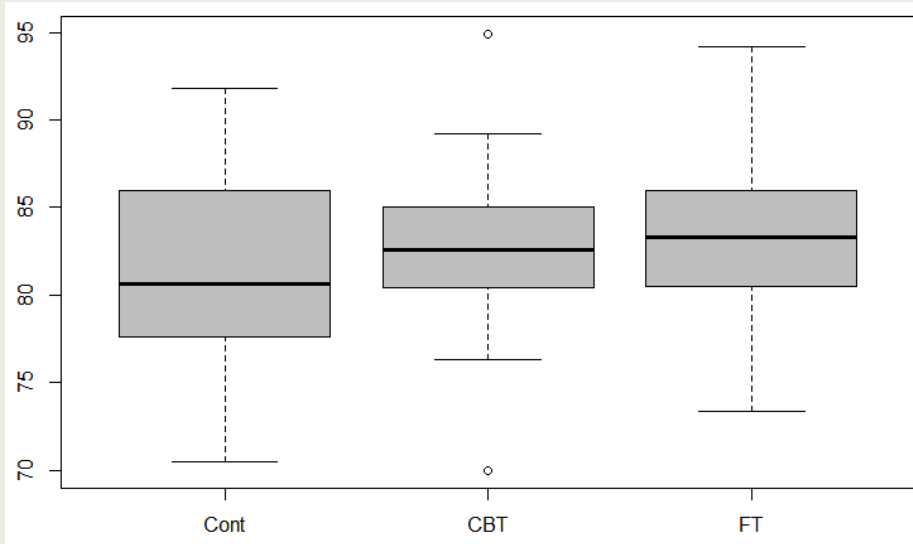
ctrl	trt1	trt2
0.5830914	0.7936757	0.4425733





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
```

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

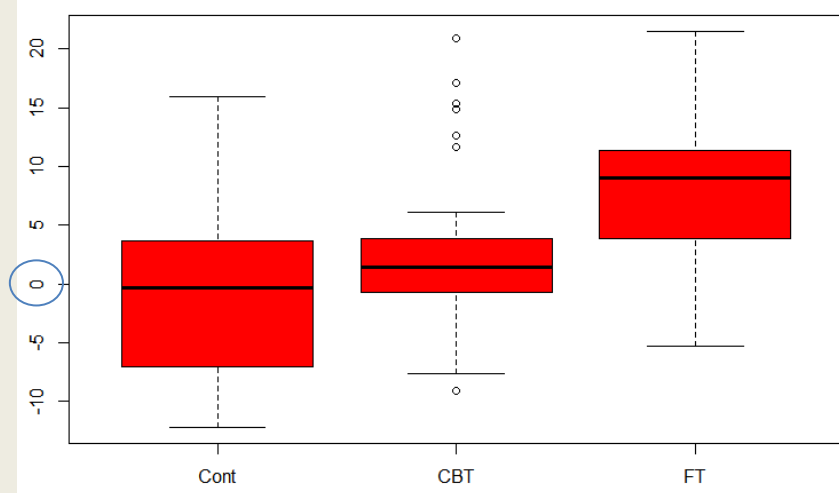
```
# 치료전 몸무게의 일원분산분석
anorexia <- read.csv("anorexia.csv")
anorexia$Treat = relevel(anorexia$Treat, ref = "cont")
levels(anorexia$Treat)
boxplot(Prewt~Treat, data=anorexia, col = 'grey')

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




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 그룹 간의 치료전후 몸무게 차이가 있는 것으로 분석되었다.
- ▶ 정규성 검정결과 잔차의 정규성이 확인되었다.



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	<u>0.16654</u>
FT - Cont == 0	7.715	2.348	3.285	<u>0.00313</u> **

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

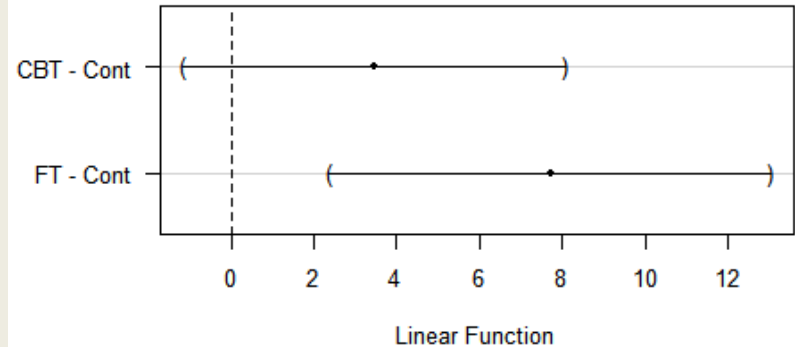
Fit: `lm(formula = diff ~ Treat, data = anorexia)`

Linear Hypotheses:

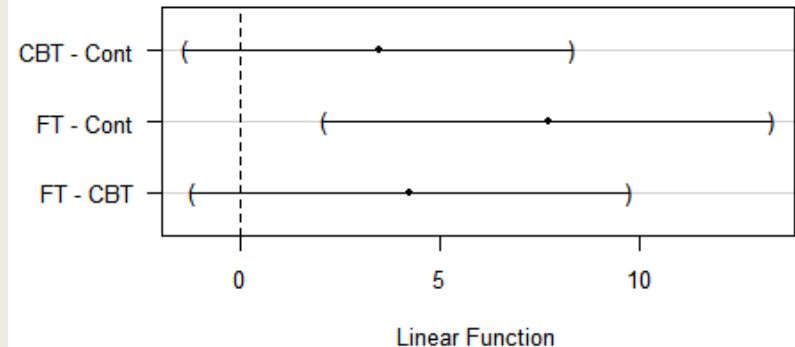
	Estimate	Std. Error	t value	Pr(> t)
CBT - Cont == 0	3.457	2.033	1.700	<u>0.21161</u>
FT - Cont == 0	7.715	2.348	3.285	<u>0.00443</u> **
FT - CBT == 0	4.258	2.300	1.852	<u>0.16005</u>

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

95% family-wise confidence level



95% family-wise confidence level



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