



통계 기반 데이터 분석

분산분석



한국기술교육대학교
온라인평생교육원



학습내용

- > 분산분석의 개념
- > 분산분석의 특성
- > R을 이용한 분산분석



학습목표

- > 분산분석의 개념에 대해 이해하고, 분석 방법을 파악할 수 있다.
- > 분산분석의 특성에 대해 이해하고 적합한 검정 방법을 채택할 수 있다.
- > 빅데이터 분석 도구인 R을 활용하여 분산분석을 시행 할 수 있다.

분산분석의 개념

왜 부부싸움이 일어날까. 성격이 다르기 때문일까? 환경적 영향, 분모의 영향, 다른 가치관을 가진 인격체로 변할 가능성도 있음. - 두 집단의 상관관계 파악 - 분산분석

두 집단의 모집단을 비교 분석하고자 할 때 - 남녀간의 의견에 대한 평균값을 비교분석 - 남녀의 시각 차이 - 독립변수에 따라 달라짐.

- 두 집단 간 속성에 대한 평균 차이를 검증하는 방법으로 사용하는 t검정의 비효율성을 줄이기 위해 집단간 변화량과 집단내 변화량을 비교하는 방법으로 사용하는 F 분포에 근거하여 검정

t검정 - 두 집단 간 속성에 대한 평균 차이를 검증하는 방법

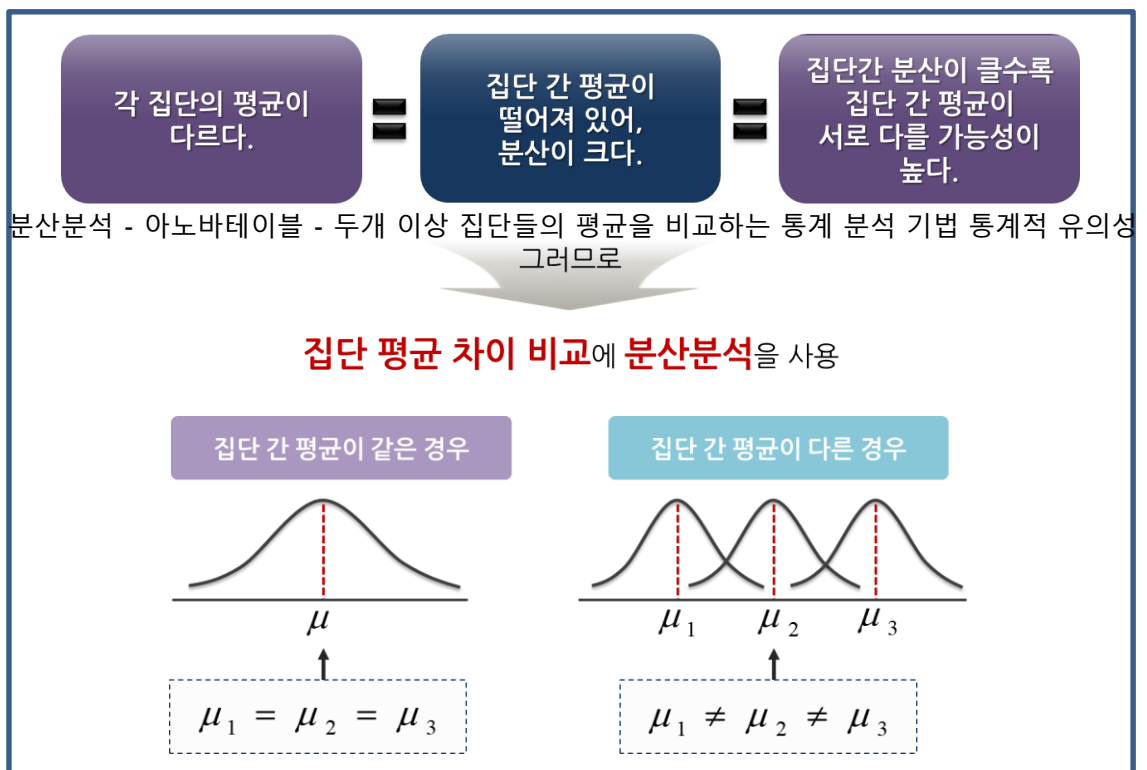
2. 분산분석

#비효율성 - 3개 이상 모집단을 비교할 때, 두 독립집단끼리 비교하는 t검정을 세번 시행하는 경우

- 두 개 이상 집단들의 평균을 비교하는 통계분석 기법
- 두 개 이상 집단들의 평균 간 차이에 대한 통계적 유의성을 검증하는 방법
- 관측자료가 몇 개의 그룹으로 구분된 경우 그룹 평균 간 차이를 그룹 내 변동에 비교하여 살펴보는 데이터 분석 방법

3. 분산분석을 사용하는 이유 (분산분석을 발전시킨, 로널드 피셔, 1890~1962)

- 집단들의 평균 차이 비교





분산분석의 특성

1. 분산분석의 기본 가정

- 정규성 : 각 집단에 해당되는 모집단의 분포가 정규분포임 분산동일
- 성=등분산성) : 각 집단에 해당되는 모집단의 분산은 모두 동일함
- 독립성 : 표본은 각 모집단에서 독립적으로(무작위로) 추출됨

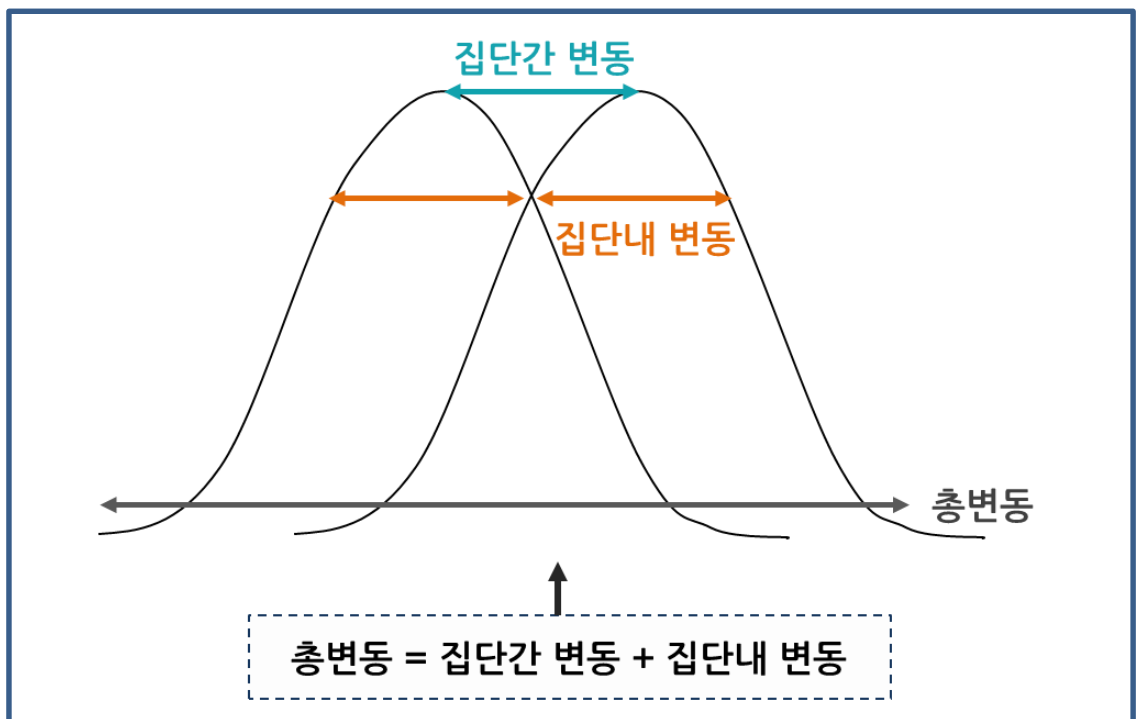
2. F통계량(F-value)

- 집단간 분산과 집단내 분산의 비

계산식

$$F = \frac{\text{집단간 분산}}{\text{집단내 분산}}$$

- 집단간 분산이 클수록, 집단내 분산이 작을수록 집단평균이 다를 가능성 증가
- 두 종류의 분산이 갖는 값의 상대적 크기에 의해 집단 간 평균의 동일성 여부가 결정됨





분산분석의 특성

3. 분산분석의 구분

분석 방법	특징
일원(배치)분산분석 (one way ANOVA)	<ul style="list-style-type: none"> • 요인(집단을 구분하는 독립변수)이 하나인 경우 • 모집단의 수에 제한이 없음 • 각 표본의 수가 같지 않아도 됨
이원(배치)분산분석 (two way ANOVA)	<ul style="list-style-type: none"> • 요인(집단을 구분하는 독립변수)이 둘인 경우 • 요인이 2개 이상인 경우, 요인이 결과에 미치는 영향을 알아보기 위한 주효과와 상호작용 효과를 살펴볼 수 있음
다원배치 분산분석 (multiple way ANOVA)	독립변수가 둘 이상인 경우를 총칭

(귀무가설 - 모집단의 평균은 모두 동일하다. , 대립가설- 적어도 두개의평균들간의 차이가 있다.

4. 분산분석의 가설 설정

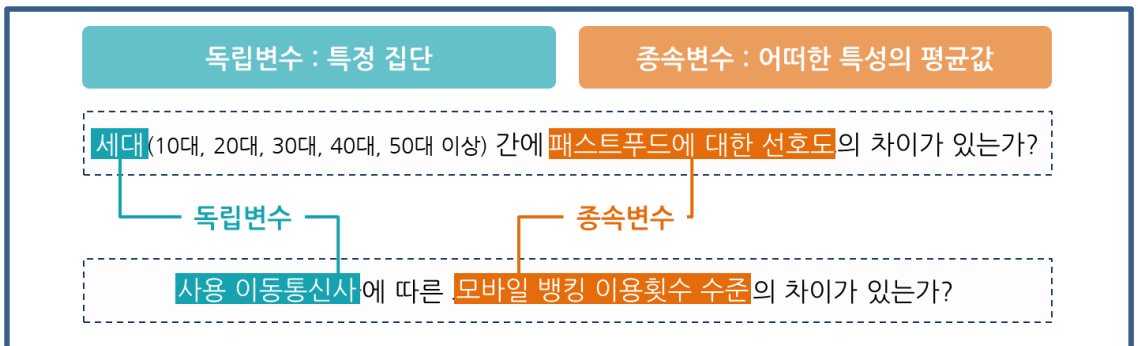
구분	H_0 (귀무가설)	H_1 (대립가설)
일원분산분석	$\mu_1 = \mu_2 = \mu_3$ (모집단평균은 모두 동일함)	적어도 두 개의 평균들 간에는 차이가 있음
이원분산분석	$\mu_1 = \mu_2 = \mu_3 = \dots \mu_n$ (모집단평균은 모두 동일함)	적어도 두 개의 평균들 간에는 차이가 있음



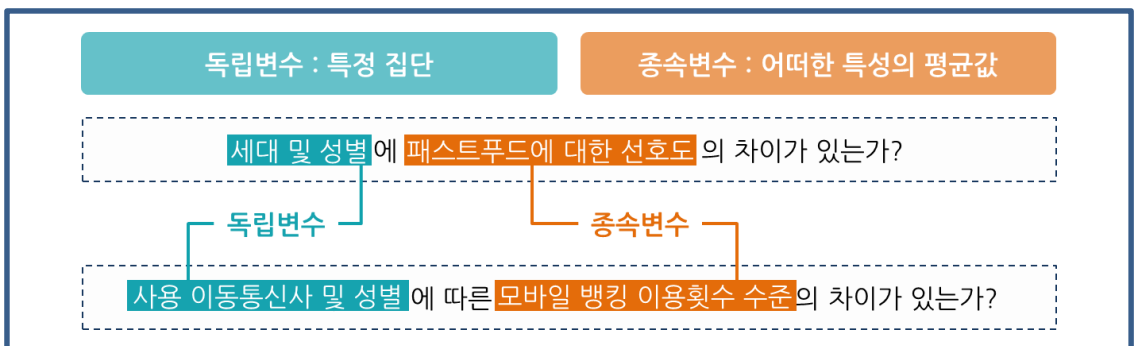
분산분석의 특성

5. 연구문제 예시

- 세대 간에 패스트푸드에 대한 선호도의 차이가 있는가?
- 사용 이동통신사에 따른 모바일 뱅킹 이용횟수 수준 차이가 있는가?
- 세대 및 성별에 따른 패스트푸드 선호도의 차이가 있는가?
- 사용 이동통신사 및 성별에 따른 모바일 뱅킹 이용횟수 수준 차이가 있는가?
- 요인 구분 : 일원분산분석-독립변수가 한 개인 경우



- 요인 구분 : 이원분산분석-집단을 구분하는 독립변수가 두 개인 경우

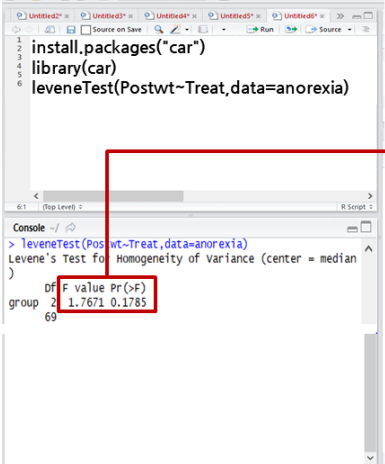


- 요인 구분 : 이원분산분석
 - 주효과 : 각 독립변수가 종속변수에 미치는 영향
 - 상호작용효과 : 여러 개의 독립변수가 상호 작용하여 나타나는 종속변수의 결과

R을 이용한 분산분석

1. R로 하는 분산분석 실습

- 분산분석은 등분산성을 가정함(검정 전 levene의 등분산 검정을 통해 등분산성을 확인)



H_0 (귀무가설)	Vs.	H_1 (대립가설)
등분산성 만족		등분산성을 만족하지 않음

유의수준 0.05이하에 귀무가설을 기각, 등분산성을 만족하는 것으로 판단하고 분산분석을 실시함

거식증 환자의 치료방법에 따른 몸무게의 변화가 있는가

- 데이터 입력(R제공 anorexia 데이터 사용) MASS의 패키지

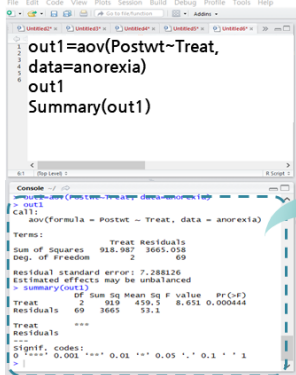
```
data(anorexia,package="MASS")
annorexia
```

R을 이용한 분산분석

1. R로 하는 분산분석 실습

■ 분산분석의 시행

2 분산분석의 시행, aov=out1 함수 이용

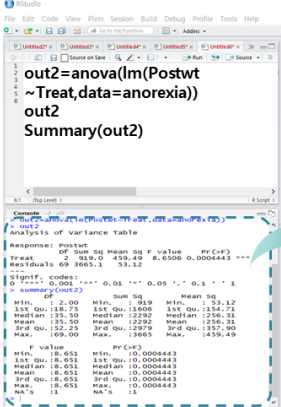


```

> out1=aov(Postwt~Treat, data=anorexia)
> out1
call:
  aov(formula=Postwt ~ Treat, data = anorexia)
Terms:
              Treat      Residuals
Sum of Squares    918.987    3665.058
Deg. of Freedom         2         69
Residual standard error: 7.288126
Estimated effects may be unbalanced
> summary(out1)
              Df Sum Sq Mean Sq F value    pr(>F)
Treat           2     919    459.5    8.651 0.000444
Residuals      69    3665     53.1
Treat
Residuals
---
Signif. Codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  
```

- P값이 0.05보다 작으므로, 귀무가설을 기각
- 각 집단의 평균은 같지 않음
- 거식증 치료 방법 별로 몸무게가 같지 않음

■ anova함수를 사용해 분산분석을 실시해 봄



```

> out2=anova(lm(Postwt
~Treat,data=anorexia))
> out2
Analysis of variance Table
Response: Postwt
              Df Sum Sq Mean Sq F value    pr(>F)
Treat           2     919.0    459.49    8.6506 0.0004443 ***
Residuals      69    3665.1     53.12
---
Signif. Codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(out2)
              Df      Sum sq      Mean sq
Min. : 2.00   Min. : 919   Min. : 53.12
1st Qu.: 18.75 1st Qu.: 1606 1st Qu.: 154.71
Median: 35.50 Median: 2292 Median: 256.31
Mean : 35.50 Mean : 2292 Mean : 256.31
3rd Qu.: 52.25 3rd Qu.: 2979 3rd Qu.: 357.90
Max. : 69.00 Max. : 3665 Max. : 459.49

              F value              Pr(>F)
Min. : 8.651   Min. : 0.0004443
1st Qu.: 8.651 1st Qu.: 0.0004443
Median: 8.651 Median: 0.0004443
Mean : 8.651 Mean : 0.0004443
3rd Qu.: 8.651 3rd Qu.: 0.0004443
Max. : 8.651 Max. : 0.0004443
NA's : 1      NA's : 1
  
```

- P값이 0.05보다 작으므로, 귀무가설을 기각
- 각 집단의 평균은 같지 않음
- 거식증 치료 방법 별로 몸무게가 같지 않음

anova는 분산분석을 하기 위한 함수이며 data는 데이터 처리를 받기위한 옵션이며 summy는평균과 pvalue를 알려주는 함수이다.

```

out2 = anova(lm(Postwt~Treat, data=anorexia))
out2
summary(out2)
  
```


⚙ R을 이용한 분산분석

1. R로 하는 분산분석 실습

- oneway함수를 사용해 분산분석을 실시해 봄

```
out3=oneway.test(Postwt~Treat,data=anorexia)
out3
Summary(out3)

> out3=oneway.test(Postwt~Treat,data=anorexia)
> out3

one-way analysis of means (not assuming
equal variances)

data: Postwt and Treat
F = 9.9047, num df = 2,000, denom df =
36.237, p-value = 0.003702

> summary(out3)
      Length Class Mode
statistic 1 -none- numeric
parameter 2 -none- numeric
p.value    1 -none- numeric
Method     1 -none- character
data.name  1 -none- character
```

```
> out3=oneway.test(postwt~Treat, data=anorexia)
> out3

One-way analysis of means(not assuming
equal variances)

data : Postwt and Treat
F = 9.9047, num df = 2,000, denom df = 36.237
p-value = 0.003702

> summary(out3)
      Length Class Mode
statistic  1 -none- numeric
parameter  2 -none- numeric
p.Value    1 -none- numeric
Method     1 -none- character
data.name  1 -none- character
>
```

- Oneway 함수는 기본적으로 등분산을 가정하지 않음
 - 등분성이 확실시 되면 “var.equal=TREU”라는 옵션을 주면 anorexia 뒤에 추가 치료제Treat 이전 이몸무게 Prewt 이후 몸무게 postwt residuals 잔차

```
out3=oneway.test(Postwt~Treat,data=anorexia)
out3
Summary(out3)

> out3=oneway.test(Postwt~Treat,data=anorexia)
> out3

one-way analysis of means (not assuming
equal variances)

data: Postwt and Treat
F = 9.9047, num df = 2,000, denom df =
36.237, p-value = 0.003702

> summary(out3)
      Length Class Mode
statistic 1 -none- numeric
parameter 2 -none- numeric
p.value    1 -none- numeric
method     1 -none- character
data.name  1 -none- character
```

```
> out3=oneway.test(postwt~Treat, data=anorexia)
> out3

One-way analysis of means(not assuming
equal variances)

data : Postwt and Treat
F = 9.9047, num df = 2,000, denom df = 36.237
p-value = 0.003702

> summary(out3)
      Length Class Mode
statistic  1 -none- numeric
parameter  2 -none- numeric
p.Value    1 -none- numeric
Method     1 -none- character
data.name  1 -none- character
>
```

- 수치는 약간 다르게 나왔지만 결과는 마찬가지로 P값이 0.05보다 작으므로, 귀무가설을 기각
- 각 집단의 평균은 같지 않음



핵심정리

1 분산분석의 개념

1. 개요

- 두 집단 간 속성에 대한 평균 차이를 검증하는 방법으로 사용하는 t검정의 비효율성을 줄이기 위해 집단 간 변화량과 집단내 변화량을 비교하는 방법으로 사용하는 F 분포에 근거하여 검정

2. 분산분석

- 두 개 이상 집단들의 평균을 비교하는 통계분석 기법

3. 분산분석을 사용하는 이유

- 집단들의 평균 차이 비교



핵심정리

2 분산분석의 특성

1. 분산분석의 기본 가정

- 정규성, 분산동일성, 독립성

2. F통계량(F-value)

- 집단간 분산과 집단내 분산의 비

3. 분산분석의 구분

분석 방법	특징
일원(배치)분산분석 (one way ANOVA)	<ul style="list-style-type: none"> 요인(집단을 구분하는 독립변수)이 하나인 경우 모집단의 수에 제한이 없음 각 표본의 수가 같지 않아도 됨
이원(배치)분산분석 (two way ANOVA)	<ul style="list-style-type: none"> 요인(집단을 구분하는 독립변수)이 둘인 경우 요인이 2개 이상인 경우, 요인이 결과에 미치는 영향을 알아보기 위한 주효과와 상호작용 효과를 살펴볼 수 있음
다원배치 분산분석 (multiple way ANOVA)	독립변수가 둘 이상인 경우를 총칭

4. 분산분석의 가설 설정

구분	H_0 (귀무가설)	H_1 (대립가설)
일원분산분석	$\mu_1 = \mu_2 = \mu_3$ (모집단평균은 모두 동일함)	적어도 두 개의 평균들 간에는 차이가 있음
이원분산분석	$\mu_1 = \mu_2 = \mu_3 = \dots \mu_n$ (모집단평균은 모두 동일함)	적어도 두 개의 평균들 간에는 차이가 있음



핵심정리

2 분산분석의 특성

5. 연구문제 예시

- 세대 간에 패스트푸드에 대한 선호도의 차이가 있는가?
- 사용 이동통신사에 따른 모바일 뱅킹 이용횟수 수준 차이가 있는가?
- 세대 및 성별에 따른 패스트푸드 선호도의 차이가 있는가?
- 사용 이동통신사 및 성별에 따른 모바일 뱅킹 이용횟수 수준 차이가 있는가?



핵심정리

3

R을 이용한 분산분석

1. R로 하는 분산분석 실습

- 거식증 환자의 치료방법에 따른 몸무게의 변화가 있는가 가설 설정
- 데이터 입력(R제공 anorexia 데이터 사용)
- anova 함수를 사용해 분산분석을 실시함
- oneway 함수를 사용해 분산분석을 실시함
- P값이 0.05보다 작으므로, 귀무가설을 기각