1. 분산분석이란?

1) 의미

분산분석(ANOVA)는 t검정의 확장이라 할 수 있다. t검정은 단 두 개의 집단 간 차이를 통계적으로 규명하는데 그치지만 분산분석은 세 개 이상의 집단 간 차이에도 적용할 수 있다. 예를 들어 귀무가설로 “세 종교 집단 간 동성연애에 대한 인식의 차이가 존재하지 않는다.”, 대립가설로 “집단 간 동성연애에 대한 인식의 차이가 존재한다.”등 으로 설정 할 수 있다. 분산분석에서 집단의 구분은 요인(factor)으로 부르는데, 분산분석은 요인에 따라 종속변수의 값이 달라지는지를 분석하는 것이다. 요인은 독립변수 역할을 한다.

2) 유형

다음의 분산분석의 유형이다.

분산분석	요인의 수	변수의 수
일원분산분석 (one-way ANOVA)	1	1
이원분산분석 (two-way ANOVA)	2	1

분산분석의 유형은 요인의 수와 종속변수의 수에 의해 결정된다.

2. 분산분석의 주요 내용

1) 분산분석의 가정

분산분석은 세 개 이상의 집단의 평균이 통계적으로 유의미한지를 분석하는 것이다. 따라서 독립변수는 명목변수로서 특정 집단에 속하는지의

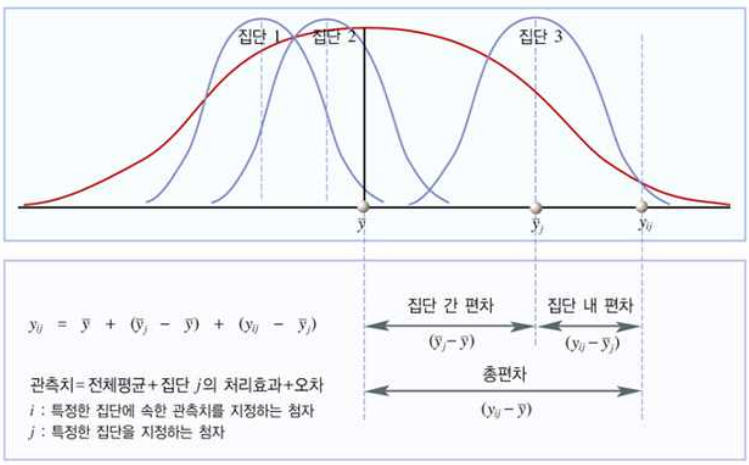
여부이며, 종속변수는 비율척도로 측정하는 것이 원칙이다.

이런 분산분석은 다음과 같은 기본 가정을 충족하여야 한다,

1. 각 표본 집단을 추출한 모집단은 동일한 분산을 가져야한다. => 등분산 가정
2. 각 표본 집단을 추출한 모집단은 정규분포여야 한다. => 정규성 가정
3. 각 표본 집단을 구성하는 관측값은 무작위로 얻은 랜덤 변수이며 서로 독립적이여야 한다. => 독립성 가성

2) 분산분석의 기본원리와 검정통계량

분산분석은 집단의 분산을 활용하여 평균의 차이를 비교하는 것이다. 예를 들어, 행정학과, 법학과 그리고 경제학과의 평균학점을 비교한다고 가정하자. 등분산 가정에 의해 평균 학점에 관한 분산이 모두 동일하고, 정규성가정에 의해 세 집단이 정규분포를 따라야 한다. 무작위로 뽑힌 표본집단에 대해 크게 두 가지 종류의 편차를 고려할 수 있다.



위 그림에서 가운데 굵은 선이 세 집단에 속한 모든 표본의 관측값 평균이라면 “집단 간 편차”와 “집단 내 편차”를 생각 할 수 있다. 집단 간 편차란 각 집단의 평균에서 전체 평균을 뺀 값이고, 집단 내 편차란 집단의

분산을 의미한다. 총 편차는 집단 간 편차와 집단 내 편차를 더한 값으로 표시된다. 따라서 분산분석 모형은 다음과 같이 정의된다.

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

여기서 y_{ij}는 j라는 집단에 속한 관측값 I를 의미하고, $\bar{y}$는 전체 평균, $(\overline{y_j} - \bar{y})$는 집단 간 편차를, $(y_{ij} - \overline{y_j})$는 집단 내 편차를 의미한다. 위의 식은 다음과 같이 바꿔 쓸 수 있다.

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

이 수식에서 $(y_{ij} - \bar{y})$의 제곱을 모두 더한 값을 총제곱합이라고 하며, SST로 표시한다. 또한 $(\overline{y_j} - \bar{y})$의 제곱을 모두 더한 값을 처리제곱합이라 하며, SSTR로 표시한다. 마지막으로 $(y_{ij} - \overline{y_j})$의 제곱을 모두 더한 값을 오차제곱합이라하고, SSE로 표시한다.

분산분석의 검정통계량은 F값이다. F값의 도출 과정을 살펴보기 위해 먼저 다음의 표를 확인해 보자

구분	제곱합	자유도	평균제곱
처리	SSTR	r-1	MSTR
오차	SSE	n-r	MSE
전체	SST	n-1	

제곱합을 자유도로 표준화한 값을 평균제곱이라고 한다. 처리의 자유도는 r-1인데 여기서 r은 분산분석 대상이 되는 모집단의 수이고 오차의 자유도는 n-r으로 n은 표본의 수이다. 이렇게 처리와 오차를 각각의 자유도로 나눈 값을 MSTR와 MSE라 한다.

분산분석의 검정통계량 F값은 다음의 수식을 통해 도출된다.

$$F = \frac{MSTR}{MSE}$$

3) 가설의 설정과 사후검정

분산분석의 귀무가설은 “집단 간 평균의 차이가 없다.”이다. 예를 들어, 우리나라 시, 군, 구 의 합계출산율 평균에 차이가 있는지를 분석하고자 한다면 귀무가설은 다음과 같다.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

예를 들어 분산분석 결과 귀무가설이 기각되어 집단 간 평균의 차이가 있다고 나타났다면 그 차이가 어디에서 발생한 차이인지 확인이 필요하다. 세 집단 모두 다를 수 있고, 시와 군은 동일하지만 구가 다를 수도 있으며, 군과 구는 동일한데 시가 다를경우도 발생할 수 있다. 따라서 대립가설은 다음과 같이 설정할 수 있다.

$$H_0 : \mu_1 \neq \mu_2 \neq \mu_3$$

$$H_0 : \mu_1 \neq \mu_2 = \mu_3$$

$$H_0 : \mu_1 = \mu_2 \neq \mu_3$$

이와 같이 어떤 집단 간에 차이가 발생하는지를 분석하는 것이 사후검정 혹은 다중비교라한다.

3. R 활용 분산분석

1) R 활용 일원분산분석

```
> setwd("C:/Users/USER/Desktop/교육원초급/data/xlsx 파일")
> library(openxlsx)
> mydata3 = read.xlsx("anova_one_way.xlsx",1)
> attach(mydata3)
> View(mydata3) # 불러온 데이터를 R Studio에서 확인
```

cities	birth_rate	ad_layer	ID
경기 수원시	1.292	자치시	1
경기 성남시	1.159	자치시	2
경기 의정부시	1.104	자치시	3
경기 안양시	1.177	자치시	4
경기 부천시	1.072	자치시	5
경기 광명시	1.235	자치시	6
경기 평택시	1.469	자치시	7
경기 동두천시	1.292	자치시	8
경기 안산시	1.219	자치시	9
경기 고양시	1.161	자치시	10

```
> names(mydata3) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"     "birth_rate" "ad_layer"   "ID"
```

독립변수는 “ad_layer”이고, 종속변수는 “birth_rate”이다.

다음은 일원분사분석의 결과이다.

```
> aov(birth_rate~ad_layer) # 종속변수 : birth_rate, 독립변수 : ad_layer
Call:
   aov(formula = birth_rate ~ ad_layer)

Terms:
                   ad_layer Residuals
Sum of Squares    4.721613 11.700008
Deg. of Freedom          2       223

Residual standard error: 0.2290555
Estimated effects may be unbalanced
```

F값은 (4.721613/2)/(11.700008/223)으로 약 45로 도출된다.

```
> aov1 = aov(birth_rate~ad_layer) # 분산분석 결과를 aov1에 저장
> summary(aov1) #aov1 결과 출력
             Df Sum Sq Mean Sq F value Pr(>F)
ad_layer      2  4.722  2.3608      45 <2e-16 ***
Residuals   223 11.700  0.0525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

유의수준이 거의 0에서 유의한 것으로 나타났다. 따라서 집단 간 차이가 있으므로 사후검정을 실시한다. 이 예제에서는 “Tukey”방법을 활용했다.

```
> TukeyHSD(aov1) # aov1의 사후검정을 시행
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = birth_rate ~ ad_layer)

$ad_layer
                       diff         lwr          upr       p adj
자치군-자치구   0.34704259   0.2587525   0.435332736 0.0000000
자치시-자치구   0.25300406   0.1628508   0.343157286 0.0000000
자치시-자치군 -0.09403854 -0.1803896 -0.007687515 0.0290742
```

사후검정 결과, '자치군-자치구'는 유의확률이 매우 낮고, '자치시-자치구' 역시 비슷한 수준이였다. 두 비교는 모두 유의수준 0.01에서 통계적으로 유의하였다. 그런데 '자치시-자치군'의 유의확률 역시 낮았지만 앞의 두 비교보다는 높았다. 이 비교는 유의수준 0.05에서 통계적으로 유의하다고 분석되었다.

2) R 활용 이원분산분석

이원분산분석은 독립변수의 역할을 하는 집단의 차원이 두 개 이며, 종속변수는 하나인 경우 활용할 수 있다. 예를들어, 이 전의 예에 '다자녀 출산 장려 조례'를 채택한 집단과 그렇지 않은 집단의 차원을 하나 더 포함하는 것을 가정하자. 표본은 첫 번째 집단의 차원인 시, 군, 구로도 구분할 수 있고, 동시에 조례를 채택한 시, 군, 구와 그렇지 않은 시, 군, 구로도 구분 가능하다. 따라서 집단의 차원이 두 개가 된다.

```
> mydata4 = read.xlsx("anova_two_way.xlsx",1) # 파을을 불러와 mydata4로 저
> attach(mydata4) # 불러온 데이터를 R과 연동
> View(mydata4) # 불러온 데이터를 R Studio에서 확인
```

cities	birth_rate	ad_layer	multichild	ID
경기 수원시	1.292	자치시	NO	1
경기 성남시	1.159	자치시	NO	2
경기 의정부시	1.104	자치시	NO	3
경기 안양시	1.177	자치시	NO	4
경기 부천시	1.072	자치시	NO	5
경기 광명시	1.235	자치시	NO	6
경기 평택시	1.469	자치시	NO	7
경기 동두천시	1.292	자치시	NO	8
경기 안산시	1.219	자치시	YES	9
경기 고양시	1.161	자치시	NO	10

```
> names(mydata4) # 불러온 데이터에 포함된 변수의 이름을 확인
[1] "cities"       "birth_rate"  "ad_layer"    "multichild"  "ID"
```

일원분산분석과 다른 점은 두 집단이 각각 분리되어 변수로 입력된다는 점과 두 집단 간 상호작용변수가 포함된다는 것이다.

```
> aov(birth_rate~ad_layer+multichild+ad_layer:multichild) # 분산분석 시행
Call:
   aov(formula = birth_rate ~ ad_layer + multichild + ad_layer:multichild)

Terms:
                    ad_layer multichild ad_layer:multichild Residuals
Sum of Squares   4.721613   0.108190              0.377127 11.214692
Deg. of Freedom         2          1                     2       220

Residual standard error: 0.2257784
Estimated effects may be unbalanced
> aov1 = aov(birth_rate~ad_layer+multichild+ad_layer:multichild) # aov1로 결
과저장
> summary(aov1) # aov1 결과 출력
                      Df Sum Sq Mean Sq F value Pr(>F)
ad_layer               2  4.722  2.3608  46.312 <2e-16 ***
multichild             1  0.108  0.1082   2.122 0.1466
ad_layer:multichild    2  0.377  0.1886   3.699 0.0263 *
Residuals            220 11.215  0.0510
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

변수의 분리를 위해 '+'기호가 사용되었고, 두 집단 간 상호작용 변수의 효과를 알아보기 위해 'ad_layer:multichild'가 입력되었다.

분석결과 행정단위의 차이를 설명하는 ad_layer는 매우 높은 F값과 매우 낮은 유의확률을 나타났다. 하지만 다자녀 조례 채택 여부를 보여주는 변

수인 multichild는 유의수준 0.1에서 유의하지 않았다. 또한 두 집단 간 상호작용 변수인 ad_layer:multichild'는 유의수준 0.056에서 통계적으로 유의하였다.

다음은 사후분석 결과이다.

```
> TukeyHSD(aov1) # aov1의 사후검정 시행
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = birth_rate ~ ad_layer + multichild + ad_layer:multichild)

$ad_layer
                       diff        lwr          upr       p adj
자치군-자치구  0.34704259  0.2600076  0.434077548 0.0000000
자치시-자치구  0.25300406  0.1641325  0.341875611 0.0000000
자치시-자치군 -0.09403854 -0.1791619 -0.008915135 0.0263119

$multichild
             diff          lwr         upr       p adj
YES-NO 0.0877627 -0.03114048 0.2066659 0.1471898

$`ad_layer:multichild`
                                  diff          lwr          upr       p adj
자치군:NO-자치구:NO      0.33269631  0.22321791  0.44217472 0.0000000
자치시:NO-자치구:NO      0.25969520  0.14704477  0.37234563 0.0000000
자치구:YES-자치구:NO     0.03739375 -0.26402560  0.33881310 0.9992350
자치군:YES-자치구:NO     0.68234375  0.34779759  1.01688991 0.0000002
자치시:YES-자치구:NO     0.20992708 -0.06721660  0.48707076 0.2524103
자치시:NO-자치군:NO     -0.07300111 -0.18027855  0.03427632 0.3709479
자치구:YES-자치군:NO    -0.29530256 -0.59475532  0.00415019 0.0557066
자치군:YES-자치군:NO     0.34964744  0.01687205  0.68242282 0.0331697
자치시:YES-자치군:NO    -0.12276923 -0.39777277  0.15223431 0.7938268
자치구:YES-자치시:NO    -0.22230145 -0.52292838  0.07832548 0.2778173
자치군:YES-자치시:NO     0.42264855  0.08881617  0.75648093 0.0045349
자치시:YES-자치시:NO    -0.04976812 -0.32604976  0.22651353 0.9954370
자치군:YES-자치구:YES    0.64495000  0.20951050  1.08038950 0.0004339
자치시:YES-자치구:YES    0.17253333 -0.22052525  0.56559192 0.8052735
자치시:YES-자치군:YES   -0.47241667 -0.89141852 -0.05341481 0.0170592
```

자치군이면서 조례를 채택하지 않은 지역(자치군 : NO)와 자치구이면서 역시 조례를 채택하지 않은 지역(자치구 : NO)의 차이는 통계적으로 유

의미하였다.

반면 자치구이면서 조례를 채택한 지역(자치구 : YES)와 자치구이면서 채택하지 않은 지역(자치구 : NO)의 차이는 통계적으로 유의하지 않았다.