# 분산분석의 이해 촉진을 위한 소고(안)

<초록>

## Ⅰ. 서론

통계분석을 위하여 분석 대상이 되는 모든 자료를 전부 수집한다는 것은 현실적으로 불가능한 경우가 대부분이다. 그래서 모집단에서 표본을 선택하게 되고, 모집단을 대표하는 표본을 추출하여 모집단에 대한 특성을 통계적으로 추론한다. 모집단의 특성을 추론하는 과정에서 표본의 평균, 분산, 표준편차는 중요한 측도가 된다.

모집단에서 표본을 뽑아 그 표본을 통해 모집단 분포를 결정하는 모수를 추정할 때, 표집분포(sampling distribution) 이론을 필요로 한다(박정식, 윤영선, 박래수, 2010). 표집분포는 모집단에서 일정한 크기로 뽑은 모든 표본들의 특성치를 나타내므로, 표집분포를 알게 되면 내가 뽑은 랜덤표본의 평균이 어느 위치에 있는지, 모집단의 평균과는 많이 떨어져 있는지 등을 알 수 있다. 모집단의 평균에 대한 가설검정을 위한 대표적인 표집분포로는 $Z$-분포, $t$-분포, $F$-분포 등이 있다. 흔히 $Z$-분포, $t$-분포는 두 집단의 모평균을 비교할 때, $F$-분포는 세 집단 이상의 모평균을 비교할 때 사용된다.

본 연구에서는 $t$-통계량과 $F$-통계량의 공통점과 차이점, 그리고 이들의 관계 등을 살펴보고자 한다. 이를 통해 $F$-통계량를 사용하는 분산분석의 이해를 돕기 위한 몇 가지 의견을 제시하고자 한다.

## Ⅱ. 이론적 배경

### 1. 표본의 특성을 나타내는 통계량

표본을 추출하여 이 표본의 분포특성을 수치화하기 위해 사용되는 방법을 통계량(statistic)이라고 한다. 모집단의 모수의 경우와 마찬가지로 표본의 특성 중 대표적인 것은 평균, 분산, 표준편차다.

#### 가. 평균

어떤 집단의 중심 측도로 가장 많이 사용되는 것은 평균(mean, average)이다. 평균은 관측값들의 모두 합하여 총관측수로 나눈 것이다(이외숙, 임용빈, 성내경, 소병수, 1996).

$$\bar{X} = \frac{\sum X_i}{n}$$ , ($X_i$ : 관측값, $\bar{X}$ : 평균, $n$ : 총관측수)

나. 분산과 표준편차

평균으로부터 퍼져 있는 정도는 각 관측값들의 평균으로부터의 편차(deviation), 절대값을 취한 편차들의 합을 생각할 수 있으나, 일반적으로 편차들을 각각 제곱한 후, 편차제곱(squared deviation)들의 평균값을 퍼짐성의 기본 측도로 사용한다. 이것을 분산(variance)이라 하고, 이 분산의 제곱근을 특히 표준편차(standard deviation)라고 한다. 표준편차는 관측 단위와 같은 단위가 되어, 상호 비교가 가능해진다. 분산과 표준편차를 식으로 표현하면 다음과 같다(이외숙 외, 1996).

분산 $S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$ , ($X_i$ : 관측값, $\overline{X}$ : 평균, $n$ : 총관측 수)

표준편차 $S = \sqrt{S^2}$

2. 표집분포

가. 평균의 표집분포

랜덤표본 $X_1, X_2, \cdots, X_n$이 있을 때, 표본평균은

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

이다. 평균과 분산의 가산 성질을 직접 적용하여 계산하면 표본평균의 기대와 분산, 표준편차는 다음과 같다.

평균의 표집분포의 평균 $E(\overline{X})$ 또는 $\mu_{\overline{x}} = \mu$

평균의 표집분포의 분산 $\mathrm{Var}(\overline{X})$ 또는 $\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}$

평균의 표집분포의 표준편차 $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

모집단이 정규분포가 아니더라도 $n$이 커질수록 평균의 표집분포는 정규분포에 접근한다. 대개 표본의 크기가 30 이상이면, 모집단의 분포 모양에 관계없이 평균의 표집분포는 정규분포를 이룬다.(박정식 외, 2010)

나. 분산의 표집분포

어떤 모집단이 $\sigma^2$의 분산을 가질 때, 이 모집단으로부터 크기가 동일하게 선택 가능한 모든 표본을 뽑아서 각각의 분산을 계산했을 때, 표본분산 $S^2$들은 일정한 분포를 이루게 되는데, 이것이 곧 분산의 표집분포다. 분산의 표집분포는 언제나 오른쪽 코리가 긴 모양을 갖지만, 표본의 크기 $n$이 커질수록 정규분포에 가까운 모양을 갖게 된다.(박정식 외, 2010)

분산의 표집분포의 평균은 모집단 분산과 같다. 즉,

$$E(S^2)=\sigma^2, \qquad S^2=\frac{\sum (X_i - \overline{X})^2}{n-1}$$

이므로 $S^2$은 $\sigma^2$에 대한 비편향추정량이다.

3. 정규분포로부터의 표본추출과 $\chi^2$, $t$, $F$ 분포

가. 정규분포

정규분포는 통계학에서 지배적인 역할을 한다. 정규모집단에 기초를 둔 표본분포들은 수학적으로 다루기가 상당히 쉽다. 표본으로부터 모집단의 본질을 유추하려면, 필히 표본 관측의 함수, 곧 통계량의 표본분포를 알 수 있어야 한다. 이러한 표본분포를 구하는 문제는 다른 어떤 분포보다도 정규분포를 가정하면 쉽게 구할 수 있다.

정규분포를 따르는 모집단에서 $n$개의 랜덤표본 $X_1$, $X_2$, …, $X_n$을 추출한다고 가정하자. 정규모집단으로부터의 랜덤표본에 대하여 표본평균의 분포는 정확히 정규분포이다. 즉,

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

표본평균의 분포는 미지의 모평균 μ에 관한 통계추정에 사용되므로, 위의 공식은 항상 기억하고 있어야 한다. 표준화하면,

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

관례적으로 표준정규분포를 따르는 확률변수를 $Z$로 표기하므로, 표준정규분포를 간단히 **$Z$ 분포**라 부르기도 한다.(이외숙 외, p. 180~181)

나. $t$ 분포

$t$ 분포는 정규분포에 관한 추론에서 모분산 $\sigma^2$의 값을 모르면서 모평균 μ를 추정할 때 쓰인다. $t$ 분포의 모양은 표준정규분포의 모양과 흡사하여 0을 중심으로 좌우대칭이나 양쪽 꼬리 부분이 더 두텁고 따라서 정점이 더 낮게 위치한다. 전체적으로 표준정규분포보다 더 퍼져 있는 형태이다. $t$ 분포는 표본 크기가 30 이하인 랜덤표본에서 모평균 μ를 추정할 때 쓰인다.

$t$ 곡선의 모양을 결정하는 것은 자유도(degrees of freedom)이다. 자유도는 $df$로 나타내며, 표본의 크기 $n$에서 1을 뺀 것이다. 자유도란 자료집단의 관찰값 중에서 자유롭게 선택될 수 있는 관찰값의 수를 말한다. 예를 들어 확률변수 $X$가 10개가 있는데 그 평균을 구했다고 하자. 이때 10개의 관찰값 중 일단 9개만 정해지면 평균을 알 수 있으므로 나머지 하나는 자동으로 정해짐을 알 수 있다. 그러므로 이 경우의 자유도 $df$는 10-1=9가 된다.(박정식 외, 2010)

다. $\chi^2$ 분포

$\chi^2$ 분포는 미지 모수 $\sigma^2$을 추정하는 데 중심적인 역할을 하는 확률밀도함수이다. 미지 모수 $\mu$를 추정할 때 표본평균을 사용했듯이 미지 모수 $\sigma^2$을 추정할 때는 표본분산

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

을 사용한다.

$X_1$, $X_2$, …, $X_n$은 평균이 $\mu$이고 분산이 $\sigma^2$인 정규분포에서의 확률표본이라 하자. 그러면

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

은 자유도가 $(n-1)$인 $\chi^2$ 분포를 따른다. 또한 $\bar{X}$와 $S^2$은 독립인 확률변수이다.(강석복 외, 2018. p.256)

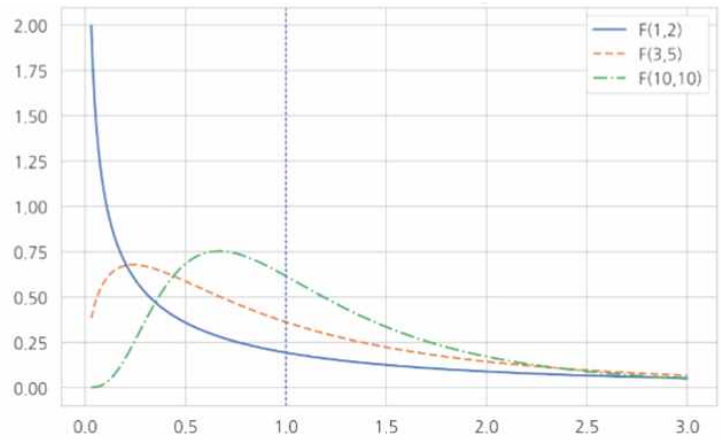
라. $\boldsymbol{F}$ 분포

$F$ 분포는 두 개 이상의 평균차를 검정하는 분산분석 방법이나 두 분산의 차이를 검정하는 경우에 적용되는 등 상당히 광범위하게 사용되는 분포이다.

$F$ 분포는 각각의 자유도로 나누어진 두 개의 $\chi^2$ 분포의 비율로 이루어지며, 다음과 같이 표시할 수 있다.

$$F(n_1-1, n_2-1) = \frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)} = \frac{S_1^2}{S_2^2}$$

값은 언제나 +기호를 갖게 되는데 그 이유는 $S_1^2$과 $S_2^2$이 모두 양수이기 때문이다. $F$-분포는 $S_1^2$의 자유도 $(n_1-1)$과 $S_2^2$의 자유도 $(n_2-1)$에 따라서 그 모양이 달라진다. 분자의 자유도와 분모의 자유도에 따라 달라지는 $F$ 분포의 모양을 몇 가지 그려보면 [그림 1]과 같다. $F$ 분포는 $n_1$, $n_2$가 크면 정규분포에 접근한다.




[그림 1] 자유도에 따른 F 분포의 모양

[https://datascienceschool.net/02%20mathematics/08.05%20%EC%8A%A4%ED%8A%9C%EB%8D%98%ED%8A%B8%20t%EB%B6%84%ED%8F%AC,%20%EC%B9%B4%EC%9D%B4%EC%A0%9C%EA%B3%B1%EB%B6%84%ED%8F%AC,%20F%EB%B6%84%ED%8F%AC.html]

$F$값은 두 분산의 비율로써 계산이 되기 때문에, $S_1^2$과 $S_2^2$이 비슷하면 $F$값은 1에 가까워진다. 그러나 두 표본의 분산으로부터 계산된 $F$값이 $F$ 분포표의 임곗값보다 매우 크다면, 이 표본들은 분산 $\sigma^2$이 서로 다른 모집단에서 뽑혔다고 할 수 있다. $F$ 분포는 위에서 말한 바와 같이 두 개의 자유도에 의해 결정되는 확률분포이므로 두 모집단에서 뽑힌 표본의 크기 $n_1$과 $n_2$에 따라 가설검정의 임곗값이 달라진다.(박정식 외, 2010. p. 289~290)

4. 두 모집단 평균에 대한 가설검정

첫 번째 모집단에서 뽑힐 수 있는 표본들과 두 번째 모집단에서 뽑힐 수 있는 모든 표본들의 평균차의 표집분포를 $(\bar{X}_1 - \bar{X}_2)$의 표집분포라고 하자.

두 모집단이 서로 독립적이고, 정규분포 또는 $n_1$과 $n_2$가 크면 중심극한정리에 의해 표집분포는 정규분포를 이루게 된다.

모집단의 분산을 모르고, 대표본일 때, $(\bar{X}_1 - \bar{X}_2)$의 표집분포의 평균과 표준편차, t-통계량은 다음과 같다.

$$\text{평균} \quad \mu_d = \mu_1 - \mu_2$$

$$\text{표준편차} \quad S_d = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_d}$$

모집단의 분산을 모르고, 소표본일 때, $(\bar{X}_1 - \bar{X}_2)$의 표집분포의 평균과 표준편차, $t$-통계량은 다음과 같다.

$$\text{평균} \quad \mu_d = \mu_1 - \mu_2$$

$$\text{표준편차} \quad S_d = S_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{단, } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$t\text{-통계량} \quad t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_d}$$

## 5. 분산분석

### 가. 분산분석의 개념

분산분석(analysis of variance: *ANOVA*)은 F 분포를 이용하여 미리 정해진 오류를 유지하면서 3개 이상의 모집단 평균이 서로 같은지 여부를 검증할 수 있다. 분산분석은 독립변수를 몇 개의 수준 또는 범주로 나누고 각 범주에 따라 나누어진 집단 간의 평균차이를 검정하는 것이다. 독립변수의 수준에 따라 나누어진 각 집단의 평균 간의 차이가 통계적으로 유의한지를 검정하는 것이므로 t 검정을 확대한 것이라고 볼 수 있다. 그러나 분산분석이 t 검정과 다른 것은, t 검정은 집단들의 평균을 비교하는 반면에, 분산분석은 집단의 분산을 사용하여 비교한다는 것이다.(박정식 외, 2010. p. 300)

분산분석의 기본가정((박정식 외, 2010, p. 303)

가정 1. 각 집단에 해당되는 모집단의 분포가 정규분포다.

가정 2. 각 집단에 해당되는 모집단의 분산이 같다.

가정 3. 각 모집단 내에서의 오차나 모집단 간의 오차는 서로 독립적이다.

### 다. 일원분산분석(백장선 외, 2021. p. 464 - 468)로 교체

1) 자료의 구성

일원분산분석은 하나의 독립변수가 여러 개의 수준으로 나누어져 있고, 각 수준에 해당되는 집단에는 여러 관찰값들이 포함되어 있는 자료를 분석하는 것이다.

관찰값은 $X_{ij}$로 표시되는데 두 개의 하위부호 중에서 $i$는 한 집단 내에서의 위치를 나타내고, $j$는 몇 번째의 집단인지를 나타낸다. 예를 들어 $X_{23}$은 3번째 집단의 2번째 관찰값을 말한다. 일반적으로 말해서 $X_{ij}$는 $j$번째 집단의 $i$번째 관찰값을 나타낸다.

일원분산분석의 자료구성(박정식 외, p. 304)

| 집단 / 관찰번호 | 집단 1 | 집단 2 | … | 집단 $j$ | … | 집단 $k$ | |
|---|---|---|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | … | $X_{1j}$ | … | $X_{1k}$ | |
| 2 | $X_{21}$ | $X_{22}$ | … | $X_{2j}$ | … | $X_{2k}$ | |
| 3 | $X_{31}$ | $X_{32}$ | … | $X_{3j}$ | … | $X_{3k}$ | |
| ⋮ | ⋮ | ⋮ | … | ⋮ | … | ⋮ | |
| $i$ | $X_{i1}$ | $X_{i2}$ | … | $X_{ij}$ | … | $X_{ik}$ | |
| ⋮ | ⋮ | ⋮ | … | ⋮ | … | ⋮ | |
| $n$ | $X_{n1}$ | $X_{n2}$ | … | $X_{nj}$ | … | $X_{nk}$ | |
| | $\overline{X}_1$ | $\overline{X}_2$ | | $\overline{X}_j$ | | $\overline{X}_k$ | $\overline{X}$ |

$\overline{X}$ : 전체평균, $\overline{X}_i$ : $j$번째 집단의 평균, $\overline{X}_{ij}$ : $j$번째 집단의 $i$번째 관찰값

2) 관찰값의 모형

분산분석을 위해서는 하나의 관측값과 모집단 평균의 차이는 크게 두 가지의 평균 차이에 의해 결정된다. 하나는 그 관측값이 포함된 집단과 모집단 평균 차이의 영향을 받고, 그 관측값이 포함된 집단의 평균과 그 관측값의 차이의 영향을 받는다. 이러한 과정을 식으로 표현하면 다음과 같다.

관찰값을 $X_{ij}$라 하면 $X_{ij}$는 다음과 같은 요소로 구성되어 있다.

일원분산분석에서 관찰값의 모형

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

$\mu$ : 전체평균

$\alpha_j$ : $j$번째 집단의 영향

$\varepsilon_{ij}$ : $j$번째 집단에 있는 관찰값 $i$의 우연적 오차

실제연구에서 얻을 수 있는 통계값으로 표현하면 다음과 같다.

$$X_{ij} = \overline{X} + (\overline{X}_j - \overline{X}) + (X_{ij} - \overline{X}_j)$$

$(\overline{X}_j - \overline{X})$ : $j$번째 집단의 평균과 전체평균 간의 차이

$(X_{ij} - \bar{X}_j)$ : 각 관찰값과 각 집단평균 간의 차이

$\bar{X}$ 를 왼쪽 항으로 옮겨 다음과 같이 변형시킬 수 있다.

$$(X_{ij} - \bar{X}) = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

왼쪽 식과 오른쪽 식을 각각 제곱하여 전체 관찰 수만큼 합하면 다음과 같이 된다.

$$\sum_j \sum_i (X_{ij} - \bar{X})^2 = \sum_j \sum_i (\bar{X}_j - \bar{X})^2 + \sum_j \sum_i (X_{ij} - \bar{X}_j)^2 + 2\sum_j \sum_i (\bar{X}_j - \bar{X})(X_{ij} - \bar{X}_j)$$

오른쪽 식의 세 번째 항은

$$\begin{aligned} 2\sum_j \sum_i (\bar{X}_j - \bar{X})(X_{ij} - \bar{X}_j) &= 2\sum_j (\bar{X}_j - \bar{X}) \sum_i ((X_{ij} - \bar{X}_j)) \\ &= 2\sum_j (\bar{X}_j - \bar{X}) \cdot 0 \\ &= 0 \end{aligned}$$

이 되어,

$$\sum_j \sum_i (X_{ij} - \bar{X})^2 = \sum_j \sum_i (\bar{X}_j - \bar{X})^2 + \sum_j \sum_i (X_{ij} - \bar{X}_j)^2$$

위의 식에서 오른쪽 항의 $\sum_j \sum_i (\bar{X}_j - \bar{X})^2$을 집단간 제곱합(sum of squares between groups : $SSB$)이라 하며, $\sum_j \sum_i (X_{ij} - \bar{X}_j)^2$은 집단내 제곱합(sum of squares within groups : $SSW$), 그리고 왼쪽 항의 $\sum_j \sum_i (X_{ij} - \bar{X})^2$은 총제곱합(total sum of squares : $SST$)이라 하는데, 총제곱합은 집단간 제곱합과 집단내 제곱합으로 구성되어 있으며, 각각의 자유도는 다음과 같다.

제곱합 : $SST = SSB + SSW$

자유도 : $N-1 = (k-1) + (N-k)$

($N$ : 총관찰수, $k$ : 집단의 수)

집단간 제곱합을 자유도로 나눈 것을 집단간 평균제곱(mean squares between groups : $MSB$)이라 하며, 집단내 제곱합을 자유도로 나눈 것을 집단내 평균제곱(mean squares within group : $MSW$)이라 한다. 집단의 수를 $k$개라 하면 평균제곱을 구하는 식은 다음과 같다.

$$\text{집단간 평균제곱 } MSB = \frac{SSB}{k-1}$$

$$\text{집단내 평균제곱 } MSW = \frac{SSW}{N-k}$$

$$MSB = \frac{SSB}{k-1} = \frac{\sum_j^k \sum_i^n (\bar{X}_i - \bar{X})^2}{k - 1}, \quad MSW = \frac{SSW}{N-k} = \frac{\sum_j^k \sum_i^n (X_{ij} - \bar{X}_i)^2}{N - k}$$

분산분석은 두 종류의 분산, 즉 집단간 분산과 집단내 분산 간의 비율을 구하는 방식으로 $F$ 검정을 한다. 만일 집단간 분산이 집단내 분산에 비해 그 비율이 크면 집단에 따른 차이가 크다는 것을 의미한다. 즉 독립변수를 몇 개의 수준으로 나누어 그 차이를 알아내는 것이 의미있다는 것을 말한다.

분산분석에서 $F$ 통계량

$$F(k-1,\ N-k) = \frac{MSB}{MSW}$$

Ⅲ. 분산분석의 실제

○ 두 모집단 평균 차이 t-검정, F-검정
교사 개입 집단, 교사 미개입 집단(N=20)

○ 세 모집단 평균 차이 F-검정
미중재 집단, 교사 개입 집단, 교사 미개입 집단(N=20)

Ⅳ. 논의

1. 두 분산의 비율인 F 분포를 통해 평균들의 동질성을 검정할 수 있다.(이론적 배경)

2. 분산분석을 통해 두 모집단의 평균 차이를 검정할 수 있다.(강석복 외, p. 471-473)

(증명) 크기 $n_1 = n_2$의 독립표본에 대한 정보를 가지고 평균 $\mu_1$과 $\mu_2$이고 같은 분산 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 인 두 정규분포의 평균을 비교한다고 가정하자. 독립표본 $t$ 검정을 이용하여 분석한 이 실험을 다른 관점에서 접근할 것이다. 두 표본에 속하는 관찰값의 전체 변동은 $X_{ij}$ 가 $j$번째 표본이고 $i$번째 관찰값이고 $\bar{X}$는 모든 $n = 2n_1$ 개의 관찰의 평균일 때,

$$SST = \sum_{j=1}^{2}\sum_{i=1}^{n_1}(X_{ij} - \bar{X})^2$$

이다. 이 양은 다음과 같이 두 부분으로 분할될 수 있다.

$$\begin{aligned} SST &= \sum_{j=1}^{2}\sum_{i=1}^{n_1}(X_{ij} - \bar{X})^2 \\ &= n_1\sum_{j=1}^{2}(\bar{X}_j - \bar{X})^2 + \sum_{j=1}^{2}\sum_{i=1}^{n_1}(X_{ij} - \bar{X}_j)^2 \\ &\qquad SSB \qquad\qquad SSW \end{aligned}$$

여기서 $\bar{X}_j$ 는 $j = 1, 2$에 대하여 $i$번째 표본에 속하는 관찰의 평균이다. 양 $SSW$를 더 자세하게 살펴보자. 우리는 기저 모분산이 같고 $n_1 = n_2$라고 가정하였다.

$$S_j^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{ij} - \bar{X}_j)^2$$

일 때,

$$\begin{aligned} SSW &= \sum_{j=1}^{2}\sum_{i=1}^{n_1}(X_{ij} - \bar{X}_j)^2 = \sum_{j=1}^{2}(n_1 - 1)S_j^2 \\ &= (n_1 - 1)S_1^2 + (n_1 - 1)S_2^2 \end{aligned}$$

이다. $n_1 = n_2$인 경우에 공통분산 $\sigma^2$의 합동 추정량(pooled estimator)은

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(n_1-1)S_1^2+(n_1-1)S_2^2}{n_1+n_2-2} = \frac{SSW}{2n_1-2}$$

로 주어진다는 것을 상기한다. 우리는 편차의 총제곱합을 두 부분으로 분할하였다. 한 부분, $SSW$는 $\sigma^2$의 합동 추정량을 구하기 위하여 $2n_1$-2로 나눌 수 있다. 오직 두 개의 처리(집단)가 있고 $n_1 = n_2$이므로 다른 부분, 처리제곱합(집단내 평균제곱합)

$$SSB = n_1\sum_{j=1}^{2}(\bar{X}_j - \bar{X})^2 = \frac{n_1}{2}(\bar{X}_1 - \bar{X}_2)^2 \quad \text{(시그마를 풀어서 계산함)}$$

은 $|\bar{X}_1 - \bar{X}_2|$가 크다면 클 것이다. 따라서 $SSB$가 크면 클수록 $\mu_1$과 $\mu_2$ 사이의 차를 나타내는 증거의 비중이 더 클 것이다. $SSB$가 $\mu_1$과 $\mu_2$ 사이의 유의한 차를 나타내기 충분히 큰 것은 언제인가?

$j = 1, 2$에 대하여 $X_{ij}$가 $E(X_{ij})=\mu_j$ 이고, $\mathrm{Var}(X_{ij})=\sigma^2$인 정규분포를 한다고 가정해 왔고 $SSW/(2n_1-2)$가 $\sigma^2$의 합동 추정량과 동일하므로

$$E\left(\frac{SSW}{2n_1-2}\right)=\sigma^2$$

이고

$$\frac{SSW}{\sigma^2} = \sum_{i=1}^{n_1}\frac{(X_{i1} - \bar{X}_1)^2}{\sigma^2} + \sum_{i=1}^{n_1}\frac{(X_{i2} - \bar{X}_2)^2}{\sigma^2}$$

은 자유도 $2n_1$-2인 $\chi^2$분포를 갖게 된다.

한편,

$$E(SSB) = \sigma^2 + \frac{n_1}{2}(\mu_1 - \mu_2)^2 \text{ (강석복 외, p. 483-484)}$$

을 의미하는 결과를 유도된다. $SSB$는 만약 $\mu_1 = \mu_2$라면 $\sigma^2$을 추정하고 $\mu_1 \neq \mu_2$라면 $\sigma^2$보다 더 큰 양을 추정하는 점에 주목한다. $\mu_1 = \mu_2$라는 가설 아래서

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2\sigma^2/n_1}}$$

는 표준정규분포를 갖는다. 따라서

$$Z^2 = \left(\frac{n_1}{2}\right)\left[\frac{(\bar{X}_1 - \bar{X}_2)^2}{\sigma^2}\right] = \frac{SSB}{\sigma^2}$$

는 자유도 1인 $\chi^2$ 분포를 갖는다.

$SSB$는 오직 표본평균 $\bar{X}_1$와 $\bar{X}_2$만의 함수인 반면 $SSW$는 오직 $S_1^2$과 $S_2^2$만의 함수이다. $i = 1, 2$에 대하여 표본평균 $\bar{X}_i$와 표본분산 $S_i^2$은 독립이라는 것을 의미한다. 표본은 독립이라 가정하였으므로 $SSB$와 $SSW$는 독립 확률변수가 된다. 그러므로 $\mu_1 = \mu_2$라는 가설 아래서

$$\frac{\frac{SSB}{\sigma^2}/1}{\frac{SSW}{\sigma^2}/2n_1-2} = \frac{SSB/1}{SSW/(2n_1-2)}$$

는 분자의 자유도는 1이고 분모의 자유도는 $2n_1$-2인 $F$ 분포를 갖는다.

각각의 자유도로 나눈 제곱합을 평균제곱(mean squares)이라 한다. 이 경우에 오차와 처리에 대한 평균제곱은

$$MSB = \frac{SSB}{1} 와\ MSW = \frac{SSW}{2n_1 - 2}$$

로 주어진다. $H_0 : \mu_1 = \mu_2$ 하에서 $MSB$와 $MSW$는 둘 다 $\sigma^2$을 추정한다. 그러나 $H_0$가 거짓이고 $\mu_1 \neq \mu_2$일 때, $MSB$는 $MSW$보다 더 큰 것을 추정하고 $MSW$보다 더 커지는 경향이 있다. $H_0 : \mu_1 = \mu_2$ 대 $H_0 : \mu_1 \neq \mu_2$ 를 검정하기 위하여 검정 통계량으로

$$F = \frac{MSB}{MSW}$$

를 사용한다.
귀무가설과의 불일치는 큰 $F$ 값으로 나타낸다. 그러므로 유의수준 α인 검정에 대한 기각역은

$$F > F_\alpha$$

이다. $F$는 분자와 분모의 자유도 각각 1과 $2n_1$-2에 의존한다.

3. 두 집단일 때, $F=t^2$이다. $t$ 분포는 $F$ 분포의 특별한 경우이다. $t$ 분포로 가설검정하는 것을 $F$ 분포로 가설검정할 수 있다.

(증명) 두 개의 표본 집단이 있고 각 표본의 크기는 $n$이며, 각 집단의 표본 평균과 표준편차는 $\bar{X}_1$ , $\bar{X}_2$와 $S_1$, $S_2$라고 하자.
먼저,

$$MSW = \frac{1}{2}(S_1^2 + S_2^2)$$

이다. 다음으로 $MSB$를 구하기 위해 다음과 같은 과정을 거칠 수 있다.

$$SSB = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{2-1}$$

$$= (\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2$$

$$= [\bar{X}_1 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)]^2 + [\bar{X}_2 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)]^2 \quad (\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)이므로)$$

$$= (\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2)^2 + (\frac{1}{2}\bar{X}_2 - \frac{1}{2}\bar{X}_1)^2$$

$$= 2(\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2)^2 \text{ (실수를 제곱해주면 항상 양수이므로)}$$

$$= \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^2$$

따라서

$$MSB = nSSB$$

$$= (n/2)(\bar{X}_1 - \bar{X}_2)^2$$

그러므로

$$F = \frac{MSB}{MSW}$$

$$= \frac{(n/2)(\bar{X}_1 - \bar{X}_2)^2}{(1/2)(S_1^2 + S_2^2)}$$

$$= \frac{(\bar{X}_1 - \bar{X}_2)^2}{(S_1^2/n + S_2^2/n)}$$

$$= \left[\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1/n + S_2/n}}\right]^2$$

위의 식에서 괄호 내에 있는 값은 $t$ 값과 같다. 따라서,

$$F = t^2$$

이다.

## 참고문헌


- Primer of biostatistics 6th edition, Stanton A Glantz, McGraw-Hill Medical Publishing Division


4. $F$ 분포가 세 수준 이상일 때 사용하지만, "집단간", "집단내"라는 두 평균제곱이 $\chi^2$ 분포를 따른다는 것에 기초하고 있다.
세 수준 이상으로 나누어진 집단 간의 평균 차이를 검정하는 분산분석은 F 분포를 사용하게 된다. 따라서 세 수준 이상의 집단들을 분석하기 위한 것일지라도 F 분포를 사용하기 위해서는 단지 두 모분산을 생각하면 된다. 하나는 집단들의 모분산이고, 다른 하나는 집단들 내에 있는 관측들의 분산들을 합한 모분산이다.

5. $MSB$, $MSW$는 $\sigma^2$의 불편추정량이다.

$E(MSB) = \sigma^2$, $E(MSW) = \sigma^2$임을 보이자.

(증명)
각각의 수준에서 관찰값들은 $N(n, \sigma^2)$로부터의 랜덤표본이다. 그 랜덤표본으로부터 계산된 표본분산 $S_j^2$이 모집단의 분산 $\sigma^2$의 불편추정량이다. 즉

$$E(S_j^2) = \sigma^2$$

이다. 따라서 $k$개의 표본분산 $S_1^2, \cdots, S_k^2$의 평균 $\bar{S}^2$도 모분산 $\sigma^2$의 불평추정량이 되어

$$E(\bar{S}^2) = \sigma^2$$

이다. $k$개의 표본분산의 평균 $\bar{S}^2$을 계산하면

$$\bar{S}^2 = \sum_{j=1}^{k} \frac{S_j^2}{k}$$

이고, 각 집단의 표본분산 $S_j^2 = \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-1}$을 대입하면 $k$개의 표본분산의 평균은

$$\bar{S}^2 = \sum_{j=1}^{k} \frac{S_j^2}{k} = \frac{\sum_{j}^{k}\sum_{i}^{n}(X_{ij} - \bar{X}_j)^2}{k(n-1)} = \frac{SSW}{N-k}$$

가 되어 집단내 평균제곱인 $MSW$와 같다. 그러므로

$$E(MSW) = \sigma^2$$

이다.

6. $MSW$와 $MSB$의 항의 수는 같다.

$SSW=\sum_{j}^{k}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$이므로 $SSW$의 항의 수는 $nk$개이다. 따라서 $MSW$의 항의 수도 $nk$개이다.

$SSB=\sum_{j=1}^{k}\sum_{i=1}^{n}(\bar{X}_j - \bar{X})^2$이므로 $SSB$의 항의 수는 $nk$개이다. 따라서 $MSB$의 항의 수도 $nk$개이다.

7. 2 수준인 경우에는 집단내 평균제곱 $MSW$가 $\sigma^2$의 불편추정량인 표본합동분산 $S_p^2$와 같다.(이외숙 외, p.330)

(증명)

2 수준의 표본의 수를 각각 $n_1$, $n_2$ 라고 하면,

$$MSW = \frac{SSW}{N-J} = \frac{\sum_{j=1}^{2}\sum_{i}(X_{ij} - \bar{X}_j)^2}{N - 2} = \frac{\sum_{i=1}^{n_1}(X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{i2} - \bar{X}_2)^2}{(n_1+n_2-2)}$$

두 모평균 비교를 위한 소표본 검정에 사용되는 검정 통계량:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1-\mu_2)}{S_p^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{단, } S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

표본합동분산 $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ 단, $S_1^2 = \frac{\sum(X_{i1}-\bar{X}_1)^2}{n_1-1}$, $S_2^2 = \frac{\sum(X_{i2}-\bar{X}_2)^2}{n_2-1}$

$(n_1-1)S_1^2=\sum(X_{i1}-\bar{X}_1)^2$ 이고, $(n_2-1)S_2^2=\sum(X_{i2}-\bar{X}_2)^2$ 이므로,

$$S_p^2 = \frac{\sum(X_{i1}-\bar{X}_1)^2+\sum(X_{i2}-\bar{X}_2)^2}{n_1+n_2-2}$$

Ⅴ. 결론

참고 문헌

강석복, 이승수, 김용구, 조영석, 김성철(2018). 수리통계학, 교우
박정식, 윤영선, 박래수(2010). 현대통계학 제5판, 서울; 다산출판사.
이외숙, 임용빈, 성내경, 소병수(1996). 통계학 입문, 서울; 경문사.