실험계획과 분석

심송용(한림대학교 데이터과학스쿨)

http://jupiter.hallym.ac.kr

a개 그룹 비교-일원배치 ANOVA

제곱합의 분해

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

이므로 양변을 제곱하고 좌변 우변을 각각 합하면

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

여기서

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..})=\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})$ 로 쓸 수 있고

$\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})=0$ (편차의 합이므로) 이라서 $i=1,2,\ldots,a$

a개 그룹 비교-일원배치 ANOVA

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{y}_{..})=0$ 임을 알 수 있다. 따라서

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

를 얻는다. 한 개의 제곱합을 두 개의 제곱합으로 분해(제곱합의 분해)

위의 세 항은 각각

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$: 전체제곱합(Sum of Squares Total; SST)

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$: 오차제곱합(Sum of Squares Error; SSE)

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2$: 처리제곱합(Sum of Squares Treatment; SSTrt)

a개 그룹 비교-일원배치 ANOVA

처리제곱합에 대한 검토

만일 $H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ 가 참이라면

$$\bar{y}_{1.} \approx \bar{y}_{2.} \approx \cdots \approx \bar{y}_{a.} \approx \bar{y}_{..}$$

일 것이고

$$\bar{y}_{1.} - \bar{y}_{..} \approx \bar{y}_{2.} - \bar{y}_{..} \approx \cdots \approx \bar{y}_{a.} - \bar{y}_{..} \approx 0$$

이므로

$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2$ 는 작은 값을 갖게 되나

H_0 가 참이 아니면 μ_i 들이 차이가 나게 되고

⇨ 따라서 표본에서 얻은 $\bar{y}_{i.}$ 들도 차이가 나게 됨

⇨ $\bar{y}_{i.} - \bar{y}_{..}$ 들이 양수/음수로 큰 값이 됨.

a개 그룹 비교-일원배치 ANOVA

⇨ $\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\overline{y}_{i.}-\overline{y}_{..})^2$ 가 큰 값이 됨.

따라서 SSTrt 값이 크면 귀무가설이 거짓일 가능성이 높음.

각 제곱합의 간편식

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{..})^2=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}^2-2y_{ij}\overline{y}_{..}+\overline{y}_{..}^2)=\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2-2\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}\overline{y}_{..}+\sum_{i=1}^{a}\sum_{j=1}^{n_i}\overline{y}_{..}^2$$

에서 $y_{..}=N\overline{y}_{..}$이므로

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}\overline{y}_{..}=\overline{y}_{..}\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}=\overline{y}_{..}y_{..}=N\overline{y}_{..}^2$$ 이고

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\overline{y}_{..}^2=N\overline{y}_{..}^2$$ 이므로

a개 그룹 비교-일원배치 ANOVA

$$\mathrm{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2 - N\bar{y}_{..}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2 - \frac{y_{..}^2}{N} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2 - CT$$

이다. 여기서

$$\boxed{\mathrm{CT} = N\bar{y}_{..}^2 = \frac{y_{..}^2}{N}}$$

로 수정항(correction term)이라고 한다.

$$\mathrm{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2) = \sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}^2 - 2\sum_{i=1}^{a}\sum_{j=1}^{n_i}y_{ij}\bar{y}_{i.} + \sum_{i=1}^{a}\sum_{j=1}^{n_i}\bar{y}_{i.}^2$$

여기서 $\bar{y}_{i.} = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i} = \dfrac{y_{i.}}{n_i} \Leftrightarrow n_i\bar{y}_{i.} = y_{i.}$ 이므로

a개 그룹 비교-일원배치 ANOVA

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}\bar{y}_{i.} = \sum_{i=1}^{a}\bar{y}_{i.}\sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{a}\left(\frac{y_{i.}}{n_i}\right)(y_{i.}) = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}$$ 이고

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\bar{y}_{i.}^2 = \sum_{i=1}^{a} n_i\bar{y}_{i.}^2 = \sum_{i=1}^{a} n_i\frac{y_{i.}^2}{n_i^2} = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}$$ 이므로

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - 2\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}\bar{y}_{i.} + \sum_{i=1}^{a}\sum_{j=1}^{n_i}\bar{y}_{i.}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}$$ 이다.

SSTrt = SST - SSE 이므로

$$SSTrt = \left(\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - CT\right) - \left(\sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}\right) = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i} - CT$$

SSTrt 도 위의 SST나 SSE 와 같은 방법으로 직접 계산할 수 있다.

a개 그룹 비교-일원배치 ANOVA

제곱합 요약

$$\mathrm{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - CT$$

$$\mathrm{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i}$$

$$\mathrm{SSTrt} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i} - CT$$

실제 계산은 SST와 SSTrt를 먼저 계산하고

SSE = SST - SSTrt

로 하는 것이 편리