실험계획과 분석

심송용(한림대학교 데이터과학스쿨)

http://jupiter.hallym.ac.kr

a개 그룹 비교-일원배치 ANOVA

자유도.

$\mathrm{SST}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2$: y_{ij} 는 N 개가 독립적인데 $\bar{y}_{..}=\sum\sum y_{ij}/N$ 으로 1개의 제약이 있음

자유도 $N-1$

$\mathrm{SSE}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$ 각 i에 대해서 n_i개의 y_{ij}가 독립적인데 $\bar{y}_{i.}=\sum y_{ij}/n_i$인 1개의 제약이 있음. 각 i에 대해서 (n_i-1)의 자유도이므로 모두 합하면 $\sum(n_i-1)=N-a$ 개의 자유도

$\mathrm{SSTrt}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2$: a개의 $\bar{y}_{i.}$이 독립적인데 이들의 가중평균인 $\bar{y}_{..}=\dfrac{n_1\bar{y}_{1.}+n_2\bar{y}_{2.}+\cdots+n_a\bar{y}_{a.}}{N}$인 1개의 제약. 따라서 자유도는 $(a-1)$

a개 그룹 비교-일원배치 ANOVA

분산분석표

요인	제곱합	자유도	평균제곱	F	유의확률
처리	$\mathrm{SSTrt} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2$	$a-1$	$\mathrm{MSTrt} = \frac{\mathrm{SSTrt}}{a-1}$	$F_0 = \frac{\mathrm{MSTrt}}{\mathrm{MSE}}$	$P = \Pr[F_{a-1,N-a} > F_0]$
오차	$\mathrm{SSE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$N-a$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-a}$		
전체	$\mathrm{SST} = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	$N-1$			

만일 $F_0 > F_{a-1,N-a;\alpha} \Leftrightarrow$ 유의확률 $P < \alpha \Leftrightarrow H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$ 를 기각

여기서

- $F_{a-1;N-a}$는 자유도 $(a-1, N-a)$인 F 분포를 따르는 확률변수
- $F_{a-1;N-a;\alpha}$는 자유도 $(a-1, N-a)$인 F 분포의 100(1-α)% 백분위수

a개 그룹 비교-일원배치 ANOVA

제곱합의 기댓값($n_1 = n_2 = \cdots = n_a = n$ 을 가정)

모형이 $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ 이고 μ, τ_i는 상수, 확률변수 ϵ_{ij}는 모두 독립이며

$E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$, $\sum_{i=1}^{a} \tau_i = 0$이므로 $\sigma^2 = Var(\epsilon_{ij}) = E(\epsilon_{ij}^2) - E(\epsilon_{ij})^2 = E(\epsilon_{ij}^2)$ 이다.

또, 모든 ϵ_{ij}는 독립이므로 $0 = Cov(\epsilon_{ij}, \epsilon_{ij'}) = E(\epsilon_{ij}\epsilon_{ij'}) - E(\epsilon_{ij})E(\epsilon_{ij'}) = E(\epsilon_{ij}\epsilon_{ij'})$임을 사용하면

$$E(\sum_{j=1}^{n} \epsilon_{ij})^2 = E(\sum_{j=1}^{n} \epsilon_{ij}^2 + \sum_{j \neq j'}^{n} \sum^{n} \epsilon_{ij}\epsilon_{ij'}) = n\sigma^2$$

모든 ϵ_{ij}는 독립이므로 $0 = Cov(\epsilon_{ij}, \epsilon_{i'j'}) = E(\epsilon_{ij}\epsilon_{i'j'}) - E(\epsilon_{ij})E(\epsilon_{i'j'}) = E(\epsilon_{ij}\epsilon_{i'j'})$임을 사용하면

$$E(\sum_{i=1}^{a}\sum_{j=1}^{n} \epsilon_{ij})^2 = E(\sum_{i=1}^{a}\sum_{j=1}^{n} \epsilon_{ij}^2 + \sum_{i \neq i'}^{a} \sum^{a} \sum_{j \neq j'}^{n} \sum^{n} \epsilon_{ij}\epsilon_{i'j'}) = an\sigma^2$$

이다. 또한

$$E(y_{..}) = E(\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \epsilon_{ij})) = \sum_{i=1}^{a}\sum_{j=1}^{n} E(\mu) + n\sum_{i=1}^{a}\tau_i + \sum_{i=1}^{a}\sum_{j=1}^{n} E(\epsilon_{ij}) = an\mu$$

$$E(y_{i.}) = E(\sum_{j=1}^{n}(\mu + \tau_i + \epsilon_{ij})) = \sum_{j=1}^{n} E(\mu) + \sum_{j=1}^{n}\tau_i + \sum_{j=1}^{n} E(\epsilon_{ij}) = n\mu + n\tau_i$$

이다.

따라서

$$E(y_{ij}^2) = E(\mu + \tau_i + \epsilon_{ij})^2 = E(\mu^2 + \tau_i^2 + \epsilon_{ij}^2 + 2\mu\tau_i + 2\mu\epsilon_{ij} + 2\tau_i\epsilon_{ij})$$

$$= \mu^2 + \tau_i^2 + \sigma^2 + 2\mu\tau_i$$

이다.

$$E(y_{..}^2) = E([\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu + \tau_i + \epsilon_{ij})]^2) = E([an\mu + n\sum_{i=1}^{a}\tau_i + \sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij})]^2) = E([an\mu + \sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij})]^2)$$

$$= E(a^2n^2\mu^2 + 2an\mu\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij} + (\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij})^2) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (1)$$

이고 $E(\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij})^2 = an\sigma^2$이므로

$$= E(a^2n^2\mu^2 + 2an\mu\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij} + (\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij})^2) = a^2n^2\mu^2 + an\sigma^2$$

이다.

a개 그룹 비교-일원배치 ANOVA

같은 방법으로

$$E(y_{i.}^2) = E([\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})]^2) = E([n\mu+n\tau_i+\sum_{j=1}^{n}\epsilon_{ij})]^2)$$

$$= E(n^2\mu^2+n^2\tau_i^2+(\sum_{j=1}^{n}\epsilon_{ij})^2+2n^2\mu\tau_i+2n\mu(\sum_{j=1}^{n}\epsilon_{ij})+2n\tau_i(\sum_{j=1}^{n}\epsilon_{ij}))$$

$$= n^2\mu^2+n^2\tau_i^2+E(\sum_{j=1}^{n}\epsilon_{ij})^2+2n^2\mu\tau_i \;= n^2\mu^2+n^2\tau_i^2+n\sigma^2+2n^2\mu\tau_i$$

이다. 따라서

$$E(SSE) \;=\; E(\sum\sum y_{ij}^2-\sum\frac{y_{i.}^2}{n}) =$$

$$= an\mu^2+n\sum_{i=1}^{a}\tau_i^2+an\sigma^2+2\mu\sum_{i=1}^{a}\tau_i-(an^2\mu^2+n^2\sum_{i=1}^{a}\tau_i^2+an\sigma^2+2n^2\mu\sum_{i=1}^{a}\tau_i)/n$$

$= an\sigma^2-a\sigma^2 = a(n-1)\sigma^2$ 이다.

a개 그룹 비교-일원배치 ANOVA

따라서

$$E(MSE) = \frac{E(SSE)}{a(n-1)} = \sigma^2$$

같은 방법으로

$$E(SSTrt) = E(\sum \frac{y_{i.}^2}{n} - \frac{y_{..}^2}{an}) = (an^2\mu^2 + n^2\sum_{i=1}^{a}\tau_i^2 + an\sigma^2)/n - \frac{a^2n^2\mu^2 + an\sigma^2}{an}$$

$$= n\sum_{i=1}^{a}\tau_i^2 + a\sigma^2 - \sigma^2$$

따라서

$$E(MSTrt) = \frac{E(SSTrt)}{a-1} = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

a개 그룹 비교-일원배치 ANOVA

즉, 귀무가설 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$의 참거짓과 상관없이

$E(MSE) = \sigma^2$: 불편추정량

이며

귀무가설 $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$이 참이면

$E(MSTrt) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} = \sigma^2$으로 MSTrt의 기댓값과 MSE의 기댓값이 모두 σ^2이라 검정통계량 $F = \frac{MSTrt}{MSE}$가 1에 가까운 값이 될 것이며

$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$이 거짓이면 τ_i의 값이 0이 아닌 양수/음수가 되므로

$E(MSTrt) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1} > \sigma^2$로 MSTrt의 값이 커진다. 결과적으로 검정통계량 F의 값이 커진다.