실험계획과 분석

[4주차]

정규분포 및 파생분포

$X_1, X_2, \ldots, X_n$ 이 정규분포 $N(\mu, \sigma^2)$에서의 확률표본이라고 하자.

확률표본(random sample): 표본내의 모든 자료가 독립이고 분포가 같음. iid(independent and identically distributed)라고도 함.

표준화

$E(X) = \mu$, $Var(X) = \sigma^2 < \infty$인 임의의 확률변수 X에 대해서

$$Z = \frac{X - \mu}{\sigma}$$

를 표준화라고 한다. 표준화된 확률변수 Z는 $E(Z) = 0$, $Var(Z) = 1$이다.

$X \sim N(\mu, \sigma^2)$이면 $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$이며 기댓값이 0 분산이 1인 정규분포는 표준정규분포라고 한다.

정규분포의 선형조합은 정규분포를 갖는다.

표본평균의 분포

표본평균의 분포

$X_1, X_2, \ldots, X_n$**이** $E(X_i) = \mu$, $Var(X_i) = \sigma^2$인 확률표본이면 표본평균 $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$의 기댓값과 분산은 각각 $E(\overline{X}) = \mu$, $Var(\overline{X}) = \frac{\sigma^2}{n}$ 이다.

- $X_1, X_2, \ldots, X_n$ 이 $N(\mu, \sigma^2)$에서의 확률표본이면 표본평균은 다시 정규분포를 갖는다. 즉, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
- $\overline{X}$를 표준화하면 $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$이다.

카이제곱 분포

χ

카이제곱 분포(chi square distribution)

- 정의: $Z_1, Z_2, \ldots, Z_\nu$ 이 $N(0,1)$에서의 확률표본이면 $\chi^2_\nu = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$의 분포를 자유도 ν인 카이제곱분포라고 한다.
- 독립인 카이제곱분포 $\chi^2_{\nu_1}$과 $\chi^2_{\nu_2}$의 합은 다시 카이제곱분포이며 자유도는 각각의 자유도의 합이다. 즉, $\chi_{\nu_1} + \chi_{\nu_2} \sim \chi_{\nu_1+\nu_2}$ (Cochran 정리)
- $X_1, X_2, \ldots, X_n$ 이 $N(\mu, \sigma^2)$에서의 확률표본이면 $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum(X_i-\mu)^2}{\sigma^2} \sim \chi^2_n$ 이다.
- μ 대신 $\overline{X}$를 사용한 $\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$이다. 단, $S^2 = \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}$

카이제곱분포 → Z가 독립이고 표준정규분포이면, 표준정규분포의 제곱합의 분포가 자유도가 ν인 카이제곱 분포라고한다.

t-분포

t-분포

- **정의**: Z와 χ^2_ν이 독립이고 각각 표준정규분포와 자유도 ν인 카이제곱분포를 따르면 $T = \frac{Z}{\sqrt{\chi^2_\nu/\nu}} \sim t_\nu$ 이다. t-분포의 자유도는 카이제곱분포의 자유도에 의해 결정된다.
- $X_1, X_2, \ldots, X_n$ 이 $N(\mu, \sigma^2)$에서의 확률표본이면 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$이고 $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

이므로
$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sqrt{S^2}}$$
$$= \frac{\sqrt{n}(\overline{X}-\mu)}{S} = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

t-분포는 표준정규분포 / 루트 (카이제곱분포 / 자유도)

t분포의 자유도는 카이제곱분포의 자유도에 의해 결정된다.

F-분포

F-**분포**

- **정의**: $\chi^2_{\nu_1}$과 $\chi^2_{\nu_2}$를 독립이고 각각 자유도가 ν_1, ν_2인 카이제곱분포를 따르는 확률변수라고 하면 $F = \dfrac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} \sim F_{\nu_1,\nu_2}$ 표시.
- 정의에 의해서 $F_{\nu_1,\nu_2} = \dfrac{1}{F_{\nu_2,\nu_1}}$이다. 따라서 $F_{\nu_1,\nu_2;\alpha} = \dfrac{1}{F_{\nu_2,\nu_1;1-\alpha}}$이다. 예를 들어 $F_{8,2;0.95} = 0.0516 = \dfrac{1}{F_{8,2;0.05}} = \dfrac{1}{19.3710} = 0.0516$

두개의 독립인 카이제곱분포의 비

자유도 ν인 t분포의 제곱은 $F_{1,\nu}$과 같은 분포. $t_\nu^2 = \left(\dfrac{Z}{\sqrt{\chi^2_\nu/\nu}}\right)^2 = \dfrac{Z^2}{\chi^2_\nu/\nu} = \dfrac{\chi^2_1/1}{\chi^2_\nu/\nu} \sim F_{1,\nu}$

$F_{1,10;0.05} = 4.964603 = t^2_{10,0.025} = 2.228139^2$

[5주차]

두 그룹 비교 - 독립 2표본

두 모집단 모평균 비교

- Sample A : $y_{11}, y_{12}, \ldots, y_{1n_1} \sim N(\mu_1, \sigma_1^2)$
- Sample B : $y_{21}, y_{22}, \ldots, y_{2n_2} \sim N(\mu_2, \sigma_2^2)$
- 모든 y_{ij} 들은 독립.

인 경우

$\bar{y}_{1.} = \sum_{j=1}^{n_1} \frac{y_{1j}}{n_1}$: 첫 번째 그룹 자료의 표본평균

$\bar{y}_{2.} = \sum_{j=1}^{n_2} \frac{y_{2j}}{n_2}$: 두 번째 그룹 자료의 표본평균

참고: 실험계획 및 분산분석에서 마침표는 마침표가 있는 위치의 첨자에 대한 합을 표시

y_{ij} $\bar{y}_{1.}$는 j위치에 .이 있으니까 j에 대한 합이다.

$$\bar{y}_{1.} \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$$

$$\bar{y}_{2.} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

만일 $\sigma_1^2 = \sigma_2^2$ 일때 (등분산)

귀무가설 $H_0 : \mu_1 - \mu_2 = 0$를 검정하는 문제 (등분산일 때 독립 2표본 t-검정)

$$\bar{y}_{1.} - \bar{y}_{2.} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ 이라 하면

$$\bar{y}_{1.} - \bar{y}_{2.} \sim N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

표준화

$$\frac{\bar{y}_{1.}-\bar{y}_{2.}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

표준화(standardization)이란 확률변수 $X\sim(\mu,\sigma^2)$일 때

$$Z=\frac{X-\mu}{\sigma}$$

을 말하며 $Z\sim(0,1)$

공통분산 σ^2의 추정치

$$S_1^2=\frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1.})^2}{n_1-1}, \quad S_2^2=\frac{\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2.})^2}{n_2-1}$$

라 할 때

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

$$S_p^2=\frac{(n_1-1)}{(n_1-1)+(n_2-1)}S_1^2+\frac{(n_2-1)}{(n_1-1)+(n_2-1)}S_2^2$$ S_1과 S_2의 가중평균

$$S_p^2=\frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1.})^2+\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1.})^2}{n_1+n_2-2}$$

$$\frac{\bar{y}_{1.}-\bar{y}_{2.}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

$$\Rightarrow \frac{\bar{y}_{1.}-\bar{y}_{2.}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

[6주차]

분산분석표

분산분석표

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

제곱합의 기댓값

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

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

E(y..)

E(yi.)

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

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

. . .

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

E(y^2..)

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

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

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.

E(y^2i.)

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

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

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

E(SSE)

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

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

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

E(MSE)

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

E(SSTrt)

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

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

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

가설검정

E(MSE) = σ^2

$\sum \tau_i = 0$ 이고 모든 $\tau_i \neq 0$

τ_i는 귀무가설이 참이면 τ_i는 0이고, 참이 아니면 적어도 하나의 τ_i는0이 아님

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

$E(MSTrt) = \sigma^2 + \dfrac{n\sum_{i=1}^{a}\tau_i^2}{a-1} = \sigma^2$으로 MSTrt의 기댓값과 MSE의 기댓값이 모두 σ^2이라 검정통계량 $F = \dfrac{MSTrt}{MSE}$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

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

결과적으로

귀무가설이 참이면

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

분자의 기댓값 = σ^2

분모의 기댓값 = σ^2

둘다 1에 가까운 값이 됨

귀무가설이 참이 아니면

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

MSE = σ^2

MSTrt > σ^2

분자가 커지니까 F가 커짐

τ들이 +-로 더 많이 커질수록 MSTrt의 기댓값은 제곱의 합에 들어가 있어서 점점더 커지는

평균들의 차이가 클수록 MSTrt의 기댓값은 커진다

F값이 커지게될 가능성이 크다.