실험계획과 분석

심송용(한림대학교 데이터과학스쿨)

http://jupiter.hallym.ac.kr

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

두 모집단 비교- 독립 2표본

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

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

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

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

두 모집단 비교- 독립 2표본

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

두 모집단 비교- 독립 2표본

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

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

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

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

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

두 모집단 비교- 독립 2표본

신뢰구간

모평균 차 $\mu_1 - \mu_2$에 대한 100(1-α)% 신뢰구간

$$1-\alpha = \Pr\left[-t_{n_1+n_2-2;\alpha/2} < \frac{\bar{y}_{1.} - \bar{y}_{2.} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2;\alpha/2} \right]$$

$$= \Pr\left[-t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \bar{y}_{1.} - \bar{y}_{2.} - (\mu_1 - \mu_2) < t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \right]$$

$$= \Pr\left[(\bar{y}_{1.} - \bar{y}_{2.}) - t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{y}_{1.} - \bar{y}_{2.}) + t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \right]$$

에서

$$\mu_1 - \mu_2 = (\bar{y}_{1.} - \bar{y}_{2.}) \pm t_{n_1+n_2-2;\alpha/2} S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

두 모집단 비교- 독립 2표본

가설검정

귀무가설 $H_0 : \mu_1 - \mu_2 = 0$이 참이면

$$t_0 = \frac{\bar{y}_{1.} - \bar{y}_{2.} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{y}_{1.} - \bar{y}_{2.}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

이므로 각각의 대립가설

- $H_1 : \mu_1 - \mu_2 > 0$ 일 때 $t_0 > t_{n_1+n_2-2;\alpha}$ 이면 H_0 기각
- $H_1 : \mu_1 - \mu_2 < 0$ 일 때 $t_0 < t_{n_1+n_2-2;1-\alpha} = -t_{n_1+n_2-2;\alpha}$ 이면 H_0 기각
- $H_1 : \mu_1 - \mu_2 \neq 0$ 일 때 $|t_0| > t_{n_1+n_2-2;\alpha/2}$ 이면 H_0 기각

  ($|t_0| > t_{n_1+n_2-2;\alpha/2} \Leftrightarrow t_0 > t_{n_1+n_2-2;\alpha/2}$ 또는 $t_0 < -t_{n_1+n_2-2;\alpha/2}$)

두 그룹 비교 - 독립 2표본

또는 유의확률 P가 유의수준 α 보다 작으면 기각

- $H_1: \mu_1 - \mu_2 > 0$ 일 때 $P = \Pr[T_{n_1+n_2-2} > t_0]$
- $H_1: \mu_1 - \mu_2 < 0$ 일 때 $P = \Pr[T_{n_1+n_2-2} < t_0]$
- $H_1: \mu_1 - \mu_2 \neq 0$ 일 때 $P = 2\Pr[T_{n_1+n_2-2} > |t_0|]$

교재 1.3절 보기

남자와 여자의 시험점수 평균 차이가 있는지 확인

남자 (y_{1j})	327	291	323	284	305
여자 (y_{2j})	308	324	353	344	341

분포가정

$$y_{ij} \sim N(\mu_i, \sigma^2) \;\; i = 1,2; \;\; j = 1,2,3,4,5$$

두 모집단 비교- 독립 2표본

가설

$$H_0 : \mu_1 - \mu_2 = 0 \text{ 대 } H_1 : \mu_1 - \mu_2 \neq 0$$

통계량

$$\overline{y}_{1.} = 306,\ \ \overline{y}_{1.} = 334,\ \ s_1^2 = 360,\ \ s_2^2 = 321.5$$

$$s_p^2 = \frac{(5-1)\times 360 + (5-1)\times 321.5}{5+5-2} = 340.75$$

검정통계량

$$t_0 = \frac{\overline{y}_{1.} - \overline{y}_{2.}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{306-334}{\sqrt{340.75}\sqrt{\frac{1}{5}+\frac{1}{5}}} = -2.3983$$

검정

$$|t_0| = 2.3983 > t_{8;0.025} = 2.306 \text{ 이므로 귀무가설 기각}$$

유의확률

$$P = 2\Pr[T_{n_1+n_2-2} > |t_0|] = 2\times 0.02164571 = 0.04329143 < \alpha = 0.05$$

두 모집단 비교- 독립 2표본

신뢰구간

$$\mu_1 - \mu_2 = (\bar{y}_{1.} - \bar{y}_{2.}) \pm t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (306 - 334 \pm 2.306\sqrt{340.75}\sqrt{\frac{1}{5} + \frac{1}{5}} = -28 \pm 26.92$$

$$= (-1.08, -54.92)$$