Data Science

데이터분석

잡초 종류에 따른 쌀 수확량 비교

독립된 두 집단 모평균 비교
독립표본 t검정

학습순서

- 독립된 두 확률변수 차이
- 표본평균의 표집
- 표본평균의 표준오차
- 독립된 두 집단의 모평균 비교 : 독립표본 t검정

독립된 두 집단에서 추출한 두 표본

독립된 집단에서 각각 추출된 표본 -> 독립표본(independent samples)

집단에서 표본추출	집단	X_1	X_2
	표본	$X_{11}, X_{12}, \dots, X_{1n_1}$	$X_{21}, X_{22}, \dots, X_{2n_2}$
	표본크기	n_1	n_2
모수	모평균	μ_{X_1}	μ_{X_2}
	모분산	$\sigma^2_{X_1}$	$\sigma^2_{X_2}$
표본통계량	표본평균	$\bar{X}_1$	$\bar{X}_2$
	표본분산	$S^2_{X_1}$	$S^2_{X_2}$
표본평균 표집	표본평균 표집 모평균	μ_{X_1}	μ_{X_2}
	표본평균 표집 모분산	$\sigma^2_{X_1} / n_1$	$\sigma^2_{X_2} / n_1$

독립된 두 집단의 확률변수 차이로 새로운 확률변수 생성

새로운 확률변수 : $D = X_2 - \bar{X}_1 \quad \text{or} \quad D = X_1 - \bar{X}_2$

표본평균차이 : $\bar{D} = \bar{X}_2 - \bar{X}_1$

표본평균차이 표집의 모평균 : $\mu_{\bar{D}} = \mu_{\bar{X}_2} - \mu_{\bar{X}_1} \sim \mu_{X_2} - \mu_{X_1}$

표본평균차이 표집의 모분산 : $\sigma^2_{\bar{D}} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma^2_{X_1}}{n_1} + \frac{\sigma^2_{X_2}}{n_2}$

$$\mathrm{Var}(\bar{D}) = \mathrm{Var}(\bar{X}_2 - \bar{X}_1) = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} - 2\mathrm{Cov}(\bar{X}_2, \bar{X}_1)$$

0 $\because$ $\mathrm{Cov}(\bar{X}_2, \bar{X}_1) = \mathrm{E}\left[(\bar{X}_2 - \mu_{X_2})(\bar{X}_1 - \mu_{X_1})\right] = 0$

독립된 두 집단의 표본평균차이 표준오차 구하기

표본평균차이 표집의 모분산 : $\sigma^2_{\bar{D}} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma^2_{X_1}}{n_1} + \frac{\sigma^2_{X_2}}{n_2}$

등분산 가정 : $\sigma^2_X = \sigma^2_{X_1} = \sigma^2_{X_2} \sim S^2_p$

통합표본분산 : $S^2_p = \frac{(n_1-1)S^2_{X_1} + (n_2-1)S^2_{X_2}}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2}$

표본평균차이 표집의 모표준오차 : $\sigma_{\bar{D}} = \sqrt{\mathrm{Var}(\bar{D})} = \sqrt{\mathrm{Var}(\bar{X}_2-\bar{X}_1)} = \sqrt{\frac{S^2_p}{n_1} + \frac{S^2_p}{n_2}}$

표본평균차이 표준오차 : $\mathrm{SE}(\bar{D}) = \sigma_{\bar{D}}$

표본평균 표집

- 무한집단에서 표본을 추출
- 표집(Sampling distribution)은 집단에서 뽑을 수 있는 표본을 일정한 크기로 모두 뽑았을 때, 표본통계량을 원소로 하는 집합
  (ex. 표본평균 표집, 표본분산 표집)
- 표본평균 표집은 표본평균을 원소로 하는 집합
- 표본평균의 표집분포는 표본평균이 나타내는 분포로 표본크기가 커지면 중심극한정리에 의하여 점점 뾰족해지는 근사 정규분포(종모양)를 나타냄

무한집단

확률변수

$$X$$

여기서, 자유도는 ∞

집단크기

$$\infty$$

무한집단

$$X_1, X_2, \cdots, X_\infty$$

모평균 Estimator

$$\mu_X = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} X_i}{N}$$

모분산 Estimator

$$\sigma_X^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} (X_i - \mu_X)^2}{N}$$

모표준편차

$$\sigma_X = \sqrt{\lim_{N\to\infty} \frac{\sum_{i=1}^{\infty} (X_i - \mu_X)^2}{\infty}}$$

표본평균 표집

확률변수

$$\bar{X}$$

표집크기

$$\infty$$

표본평균 표집

$$\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_\infty$$

확률변수변환 : t_{df}분포

$$\bar{X} \to t_{df}$$

$$\frac{\bar{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \sim t_n$$

여기서, n은 표본크기

표본평균 기대값(표집의 모평균)

$$\mathrm{E}[\bar{X}] = \mu_{\bar{X}} \sim \mu_X$$

표본평균 표집의 모분산

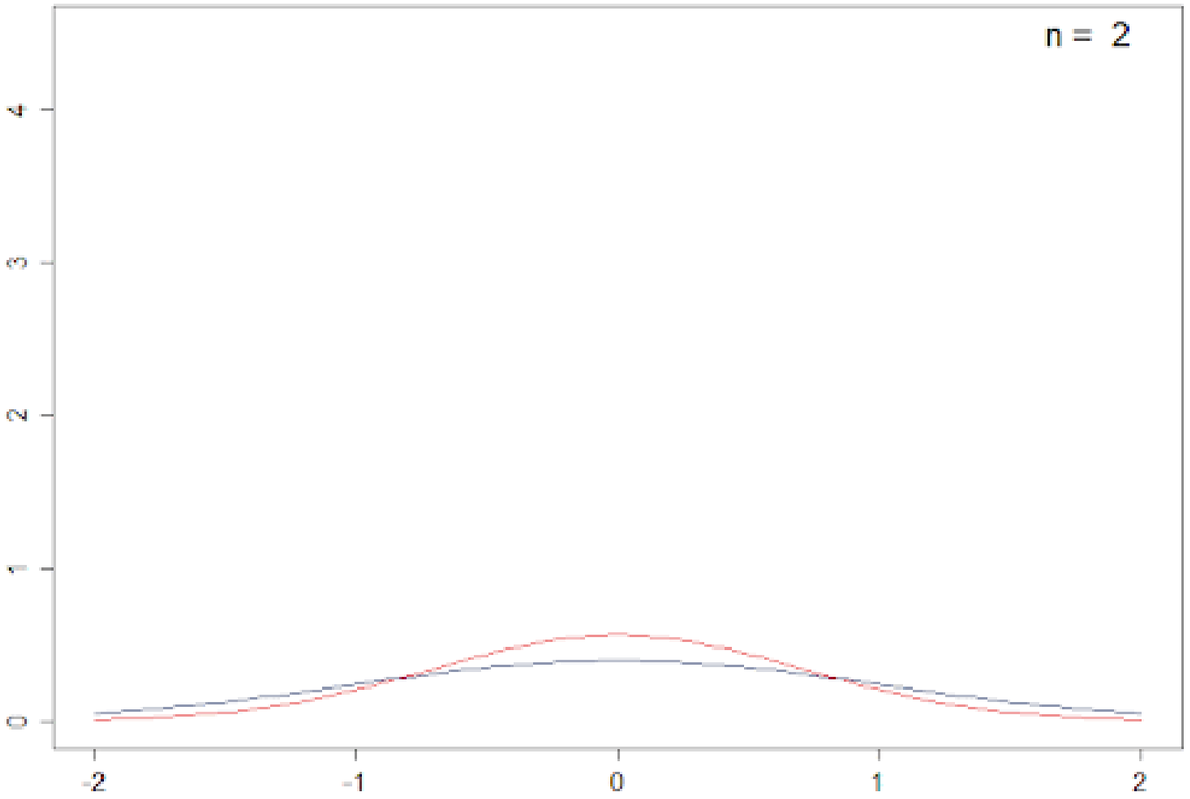
$$\mathrm{Var}(\bar{X}) = \sigma_{\bar{X}}^2 \sim \frac{\sigma_X^2}{n}$$

여기서, n은 표본크기

표본평균 표집의 모표준편차

$$\mathrm{SD}(\bar{X}) = \sigma_{\bar{X}} \sim \sqrt{\frac{\sigma_X^2}{n}}$$

표본평균 확률분포 = 표본평균 표집의 확률분포



표준정규분포를 나타내는 집단에서 추출한 표본의 표본평균이
표본크기(n)에 따라 변하는 확률밀도함수

랜덤하게 추출한 표본 : $\{X_1, \ldots, X_n\}$

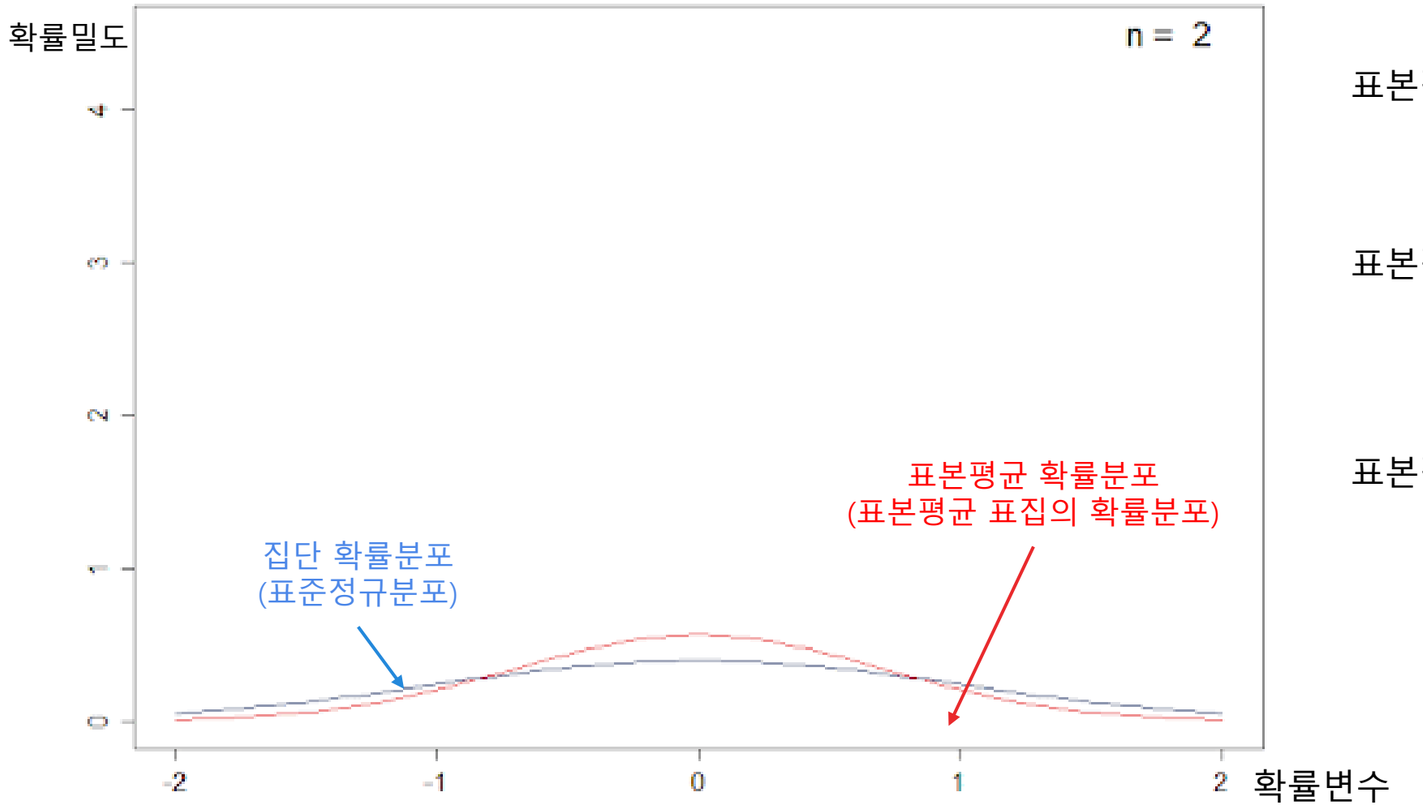
표본평균 : $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$

표본평균 표집의 평균
(표본평균 기대값) $\mathrm{E}[\bar{X}] = \mu_{\bar{X}} = \mu_X$

표본평균 표집의 분산 : $\mathrm{Var}(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$

표본평균 표집의 Z변환 : $\frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim \mathrm{N}(0,1)$

표본평균의 표준오차 = 표본평균 표집의 표준편차



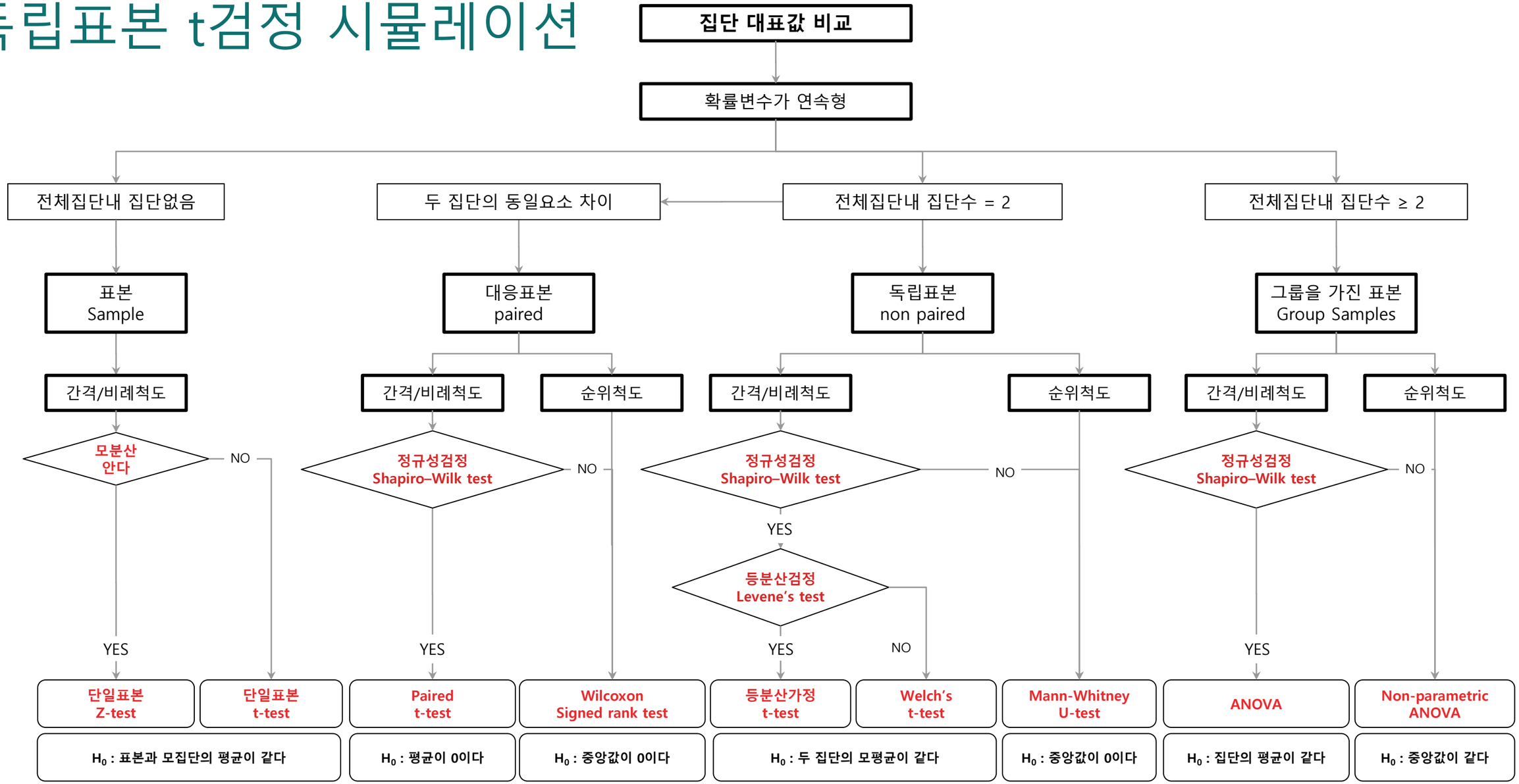
표준정규분포를 가지는 집단에서 표본의 크기를 0에서 100까지 변화시키면서 표본평균의 확률분포를 관찰

표본평균 표집의 모분산 : $\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n} \sim \frac{S^2_X}{n}$

표본평균 표집의 표준편차 : $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sim \frac{S_X}{\sqrt{n}}$

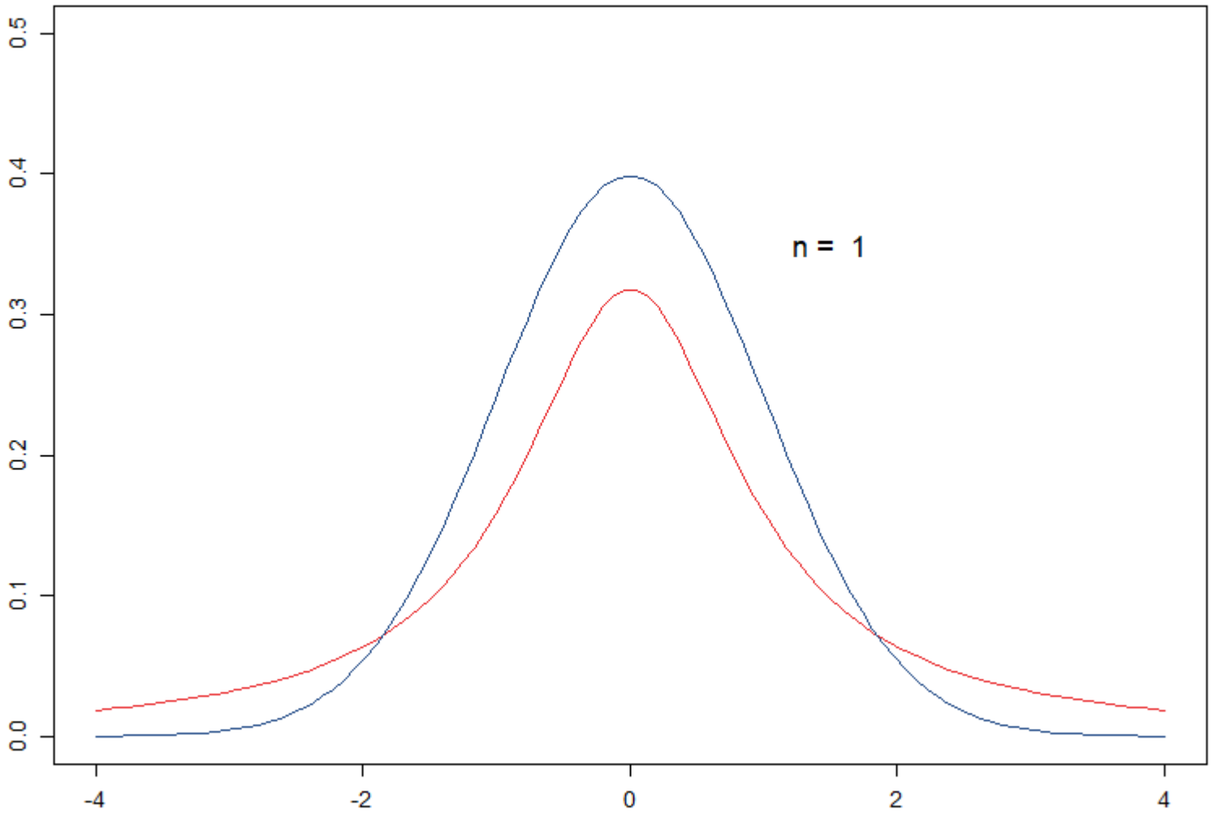
표본평균의 표준오차 : $\mathrm{SE}(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sim \frac{S_X}{\sqrt{n}}$

독립표본 t검정 시뮬레이션



표본평균차이의 표집분포를 t분포로 변환

귀무가설 : $\mu_2 - \mu_1 = 0$



자유도가 증가함에 따라 t분포가 Z분포(표준정규분포)에 수렴

$$\sigma_X^2 = \sigma_{X_1}^2 = \sigma_{X_2}^2 \sim S_p^2$$

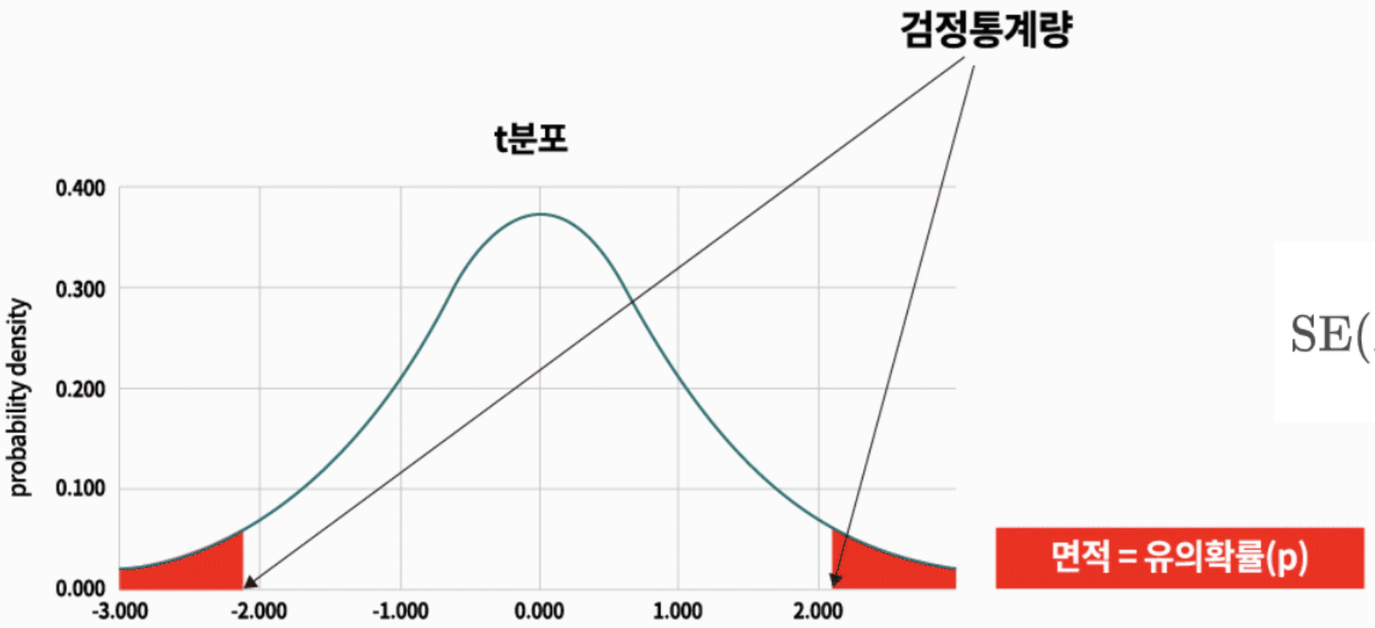
$$S_p^2 = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}$$

$$\mathrm{SE}(\bar{D}) = \sqrt{\mathrm{Var}(\bar{D})} = \sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}} \sim \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\mathrm{SE}(\bar{D})} \sim \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

t분포에서의 검정통계량과 유의확률 구하기

귀무가설 : $\mu_2 - \mu_1 = 0$

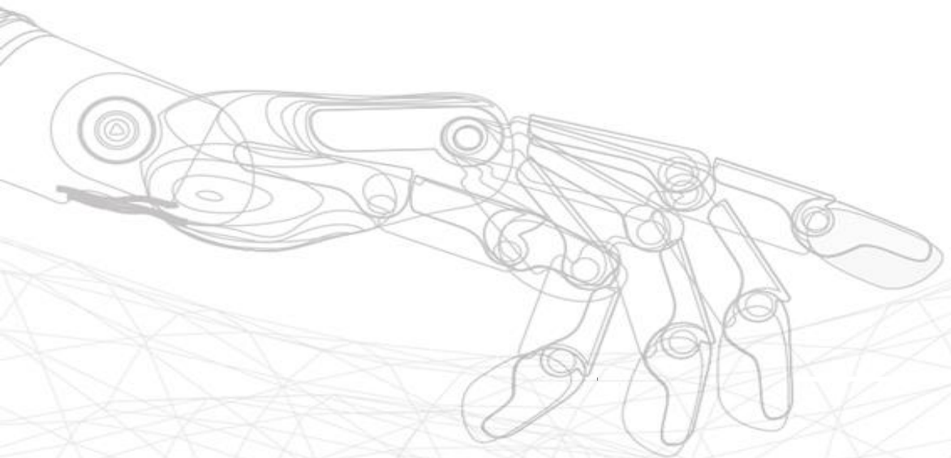


$$\sigma_X^2 = \sigma_{X_1}^2 = \sigma_{X_2}^2 \sim S_p^2$$

$$S_p^2 = \frac{(n_1 - 1)S_{X_1}^2 + (n_2 - 1)S_{X_2}^2}{n_1 + n_2 - 2}$$

$$\mathrm{SE}(\bar{D}) = \sqrt{\mathrm{Var}(\bar{D})} = \sqrt{\frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}} \sim \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\mathrm{SE}(\bar{D})} \sim \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$



감사합니다

www.datadata.link