Data Science

데이터분석

한우 등심과 설도의 지방함량 비교

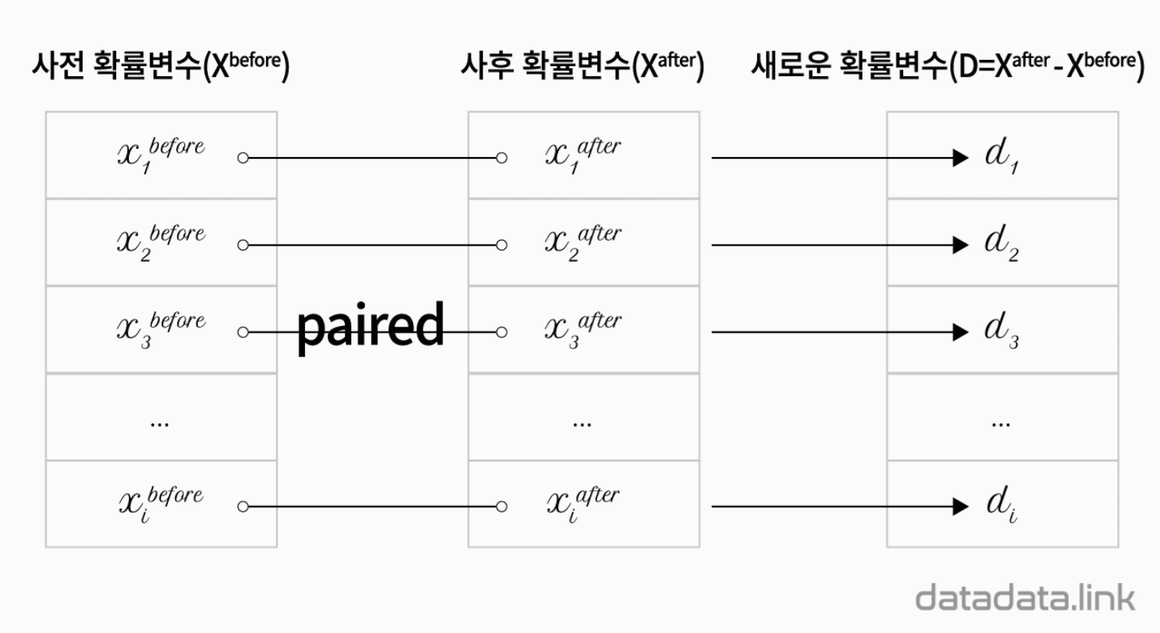
대응된 두 변수 모평균 비교

대응표본 t검정

학습순서

- 대응된 두 확률변수 차이
- 표본평균의 표집
- 표본평균의 표준오차
- 대응된 두 변수의 모평균 비교 : 대응표본 t검정

대응된(paired) 두 확률변수 차이로 새로운 확률변수 생성



대응된(paired) 두 확률변수의 차이로 새로운 확률변수 D 생성

새로운 확률변수 *D*

$$D = X_2 - X_1 \qquad d_i = x_i^{after} - x_i^{before}$$

대응표본평균

$$\overline{d_i} = \frac{\sum_{i=1}^{n} d_i}{n}$$

대응표본분산

$$S_D^2 = \frac{\sum_{i=1}^{n} (d_i - \overline{d_i})^2}{n-1}$$

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

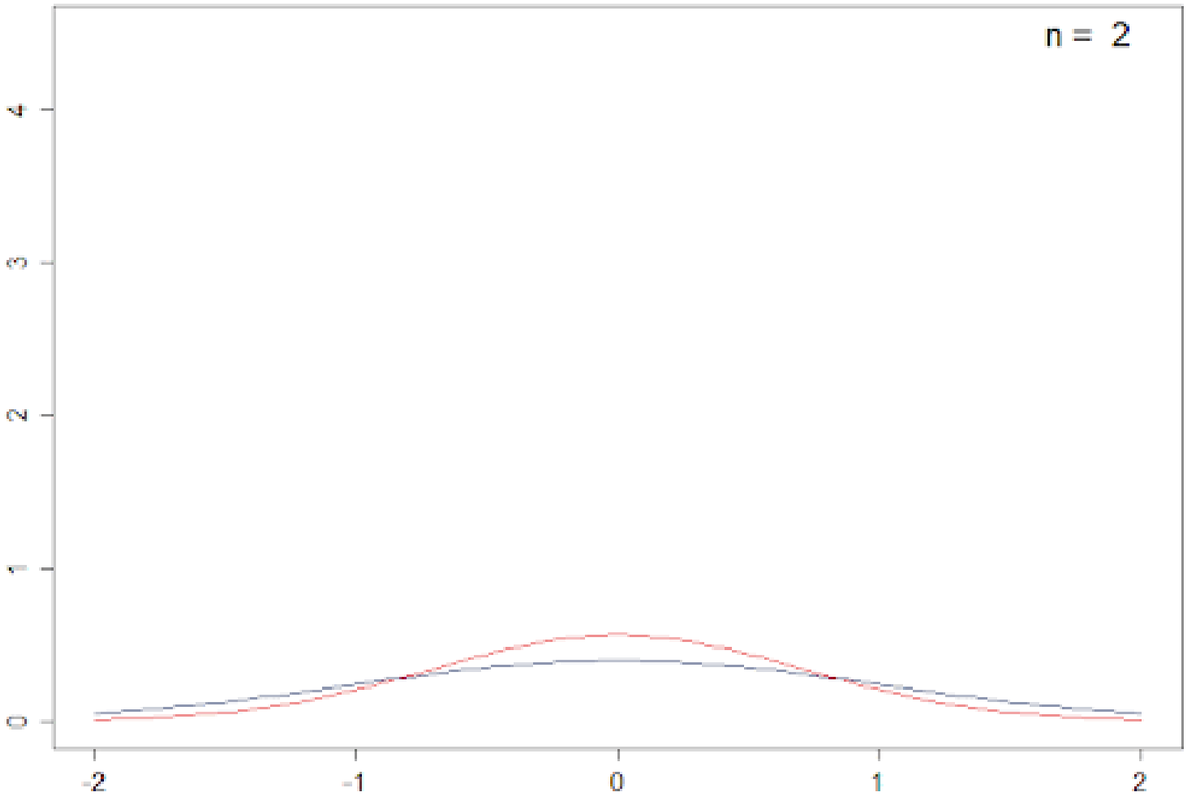
$$\mathrm{Var}(\bar{X}) = \sigma_{\bar{X}}^2 \sim \frac{\sigma_X^2}{n}$$

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 : $\{X_1, \ldots, X_n\}$

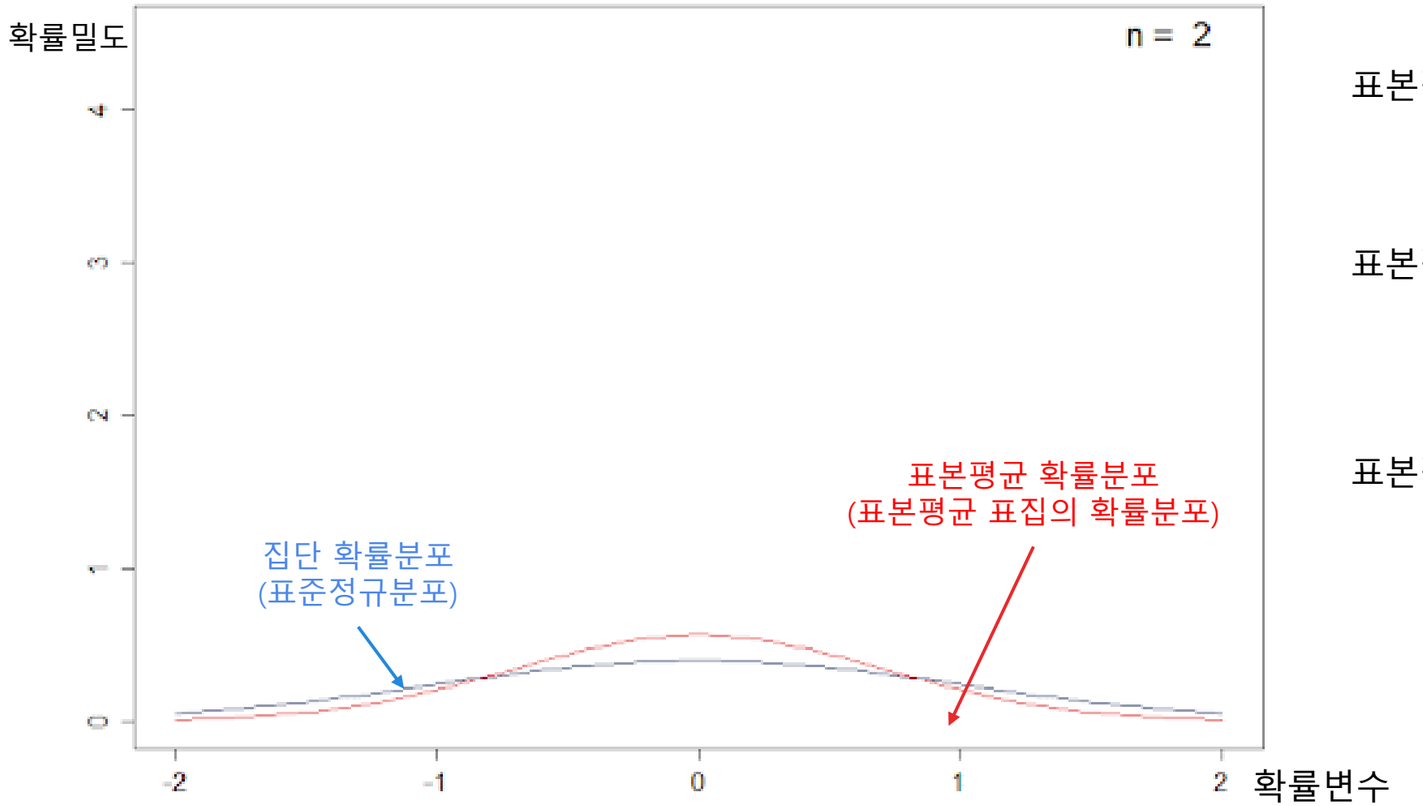
표본평균 : $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$

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



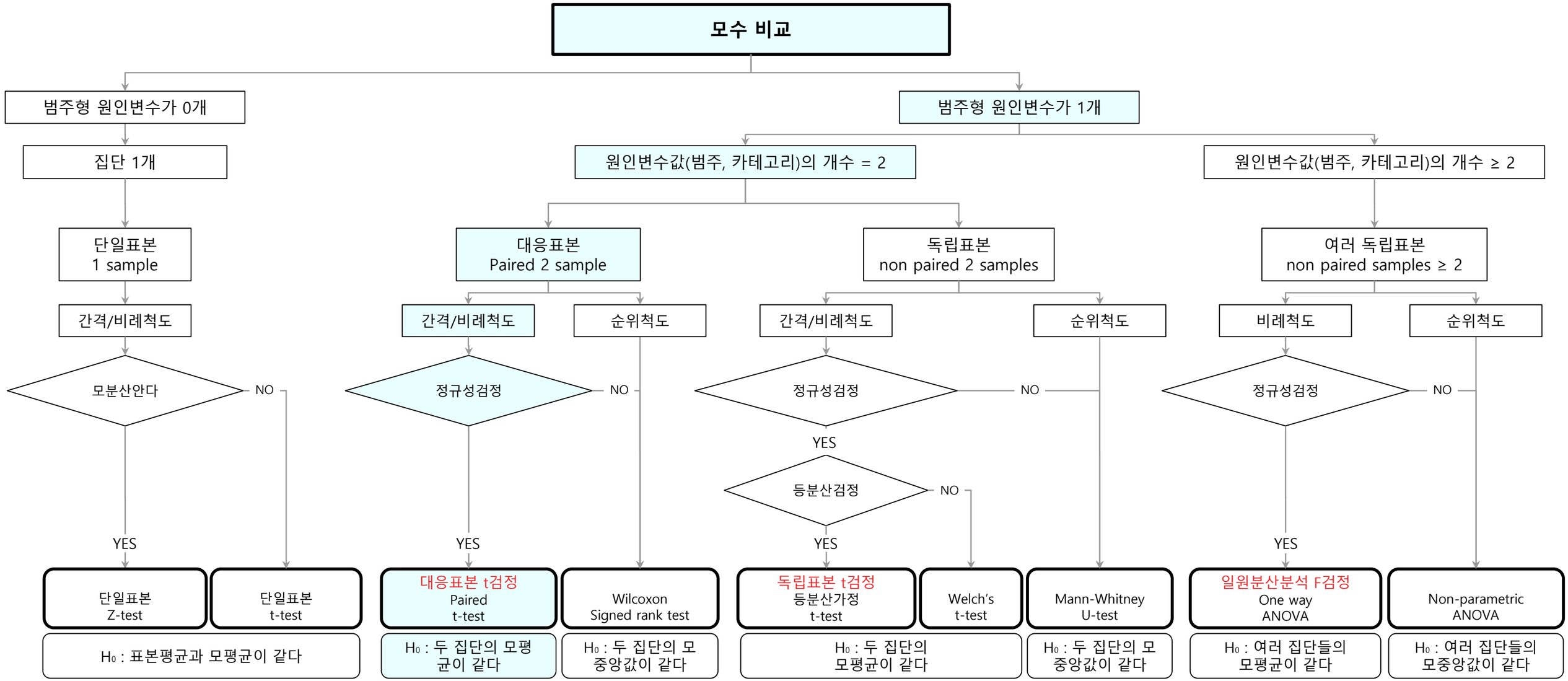
표준정규분포를 가지는 집단에서 표본의 크기를 0에서 100까지 변화시키면서 표본평균의 확률분포를 관찰

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

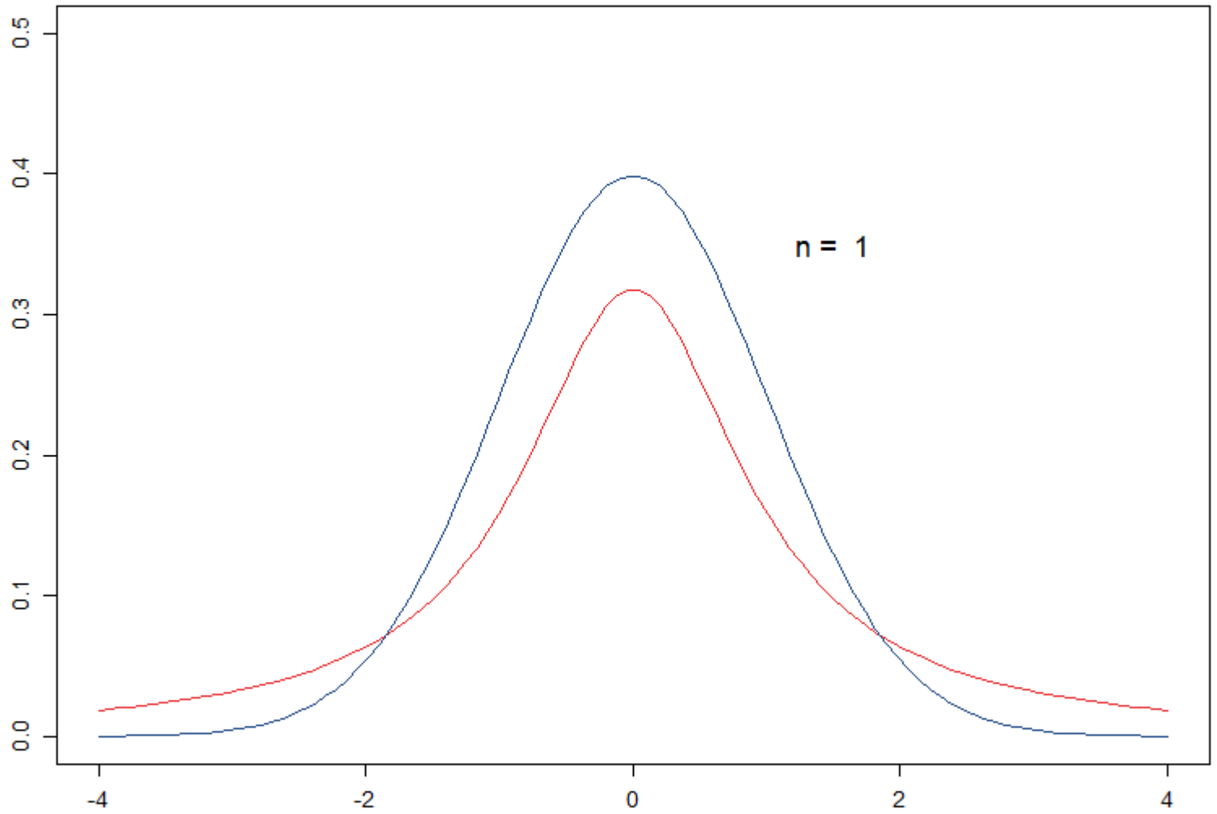
표본평균 표집의 표준편차 : $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sim \frac{S_X}{\sqrt{n}}$

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

대응표본 t검정 시뮬레이션



대응표본평균 표집의 확률분포를 t분포로 변환

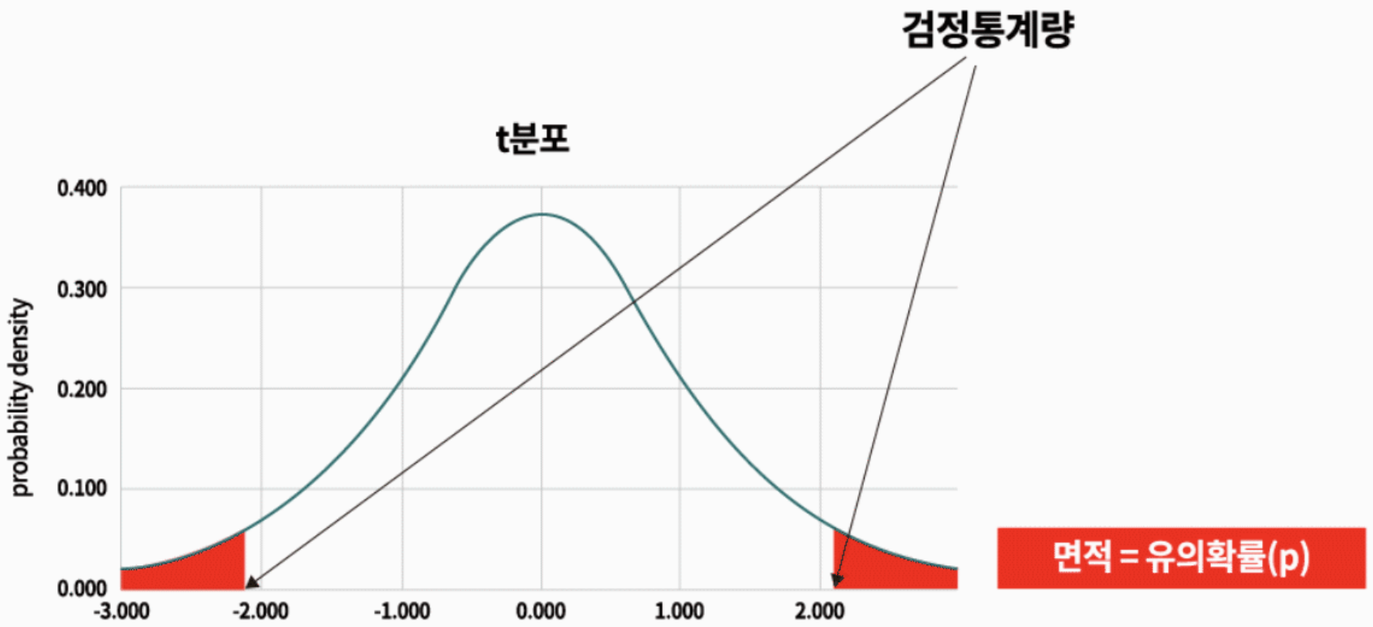


자유도가 증가함에 따라 t분포가 Z분포(표준정규분포)에 수렴

$$t = \frac{(\bar{X}_2 - \bar{X}_1) - D_0}{\frac{S_D}{\sqrt{n}}} = \frac{\bar{D} - D_0}{\frac{S_D}{\sqrt{n}}}$$

$\bar{D}$	대응표본평균
S_D	대응표본표준편차
n	대응표본크기
$n-1$	대응표본의 자유도

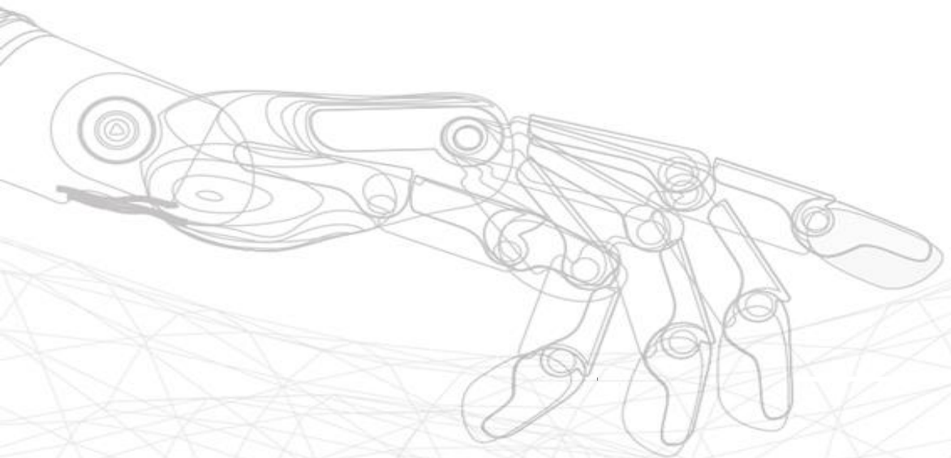
t분포에서의 검정통계량과 유의확률 구하기



$$t = \frac{(\bar{X}_2 - \bar{X}_1) - D_0}{\frac{S_D}{\sqrt{n}}} = \frac{\bar{D} - D_0}{\frac{S_D}{\sqrt{n}}}$$

$\bar{D}$	관측한 대응표본평균
D_0	귀무가설로 주어진 값
S_D	관측한 대응표본표준편차
n	대응표본크기
n-1	대응표본의 자유도

$$\mathrm{SE}(\bar{D}) = \sqrt{\mathrm{Var}(\bar{D})} = \sigma_{\bar{D}} = \sqrt{\frac{\sigma_D^2}{n}} \sim \sqrt{\frac{S_D^2}{n}} = \frac{S_D}{\sqrt{n}}$$



감사합니다

www.datadata.link