Data Science

데이터분석

잡초밀도에 따른 쌀 수확량 비교

여러 집단 모평균 비교
일원분산분석 F검정

학습순서

- 집단간분산과 집단내분산의 비
- 표본분산의 표집
- 표본분산의 표준오차
- 여러 집단 모평균 비교 : 일원분산분석 F검정

분산을 변동과 자유도로 나누어 분석

- 분산분석은 분산을 변동과 자유도로 나누어 분석하는 것
- 총변동은 처리에 의한 변동과 오차에 의한 변동으로 분할
- 총자유도는 처리의 자유도(집단간 자유도)와 오차의 자유도(집단내 자유도)로 분할

변동의 분할과 등식

$$\text{총제곱합}(SS_T) = \text{처리제곱합}(SS_{Tr}) + \text{오차제곱합}(SS_E)$$

$$\sum_{i=1}^{k}\sum_{j=1}^{k}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}(\overline{Y_i}-\bar{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}(Y_{ij}-\overline{Y_i})^2$$

자유도의 분할과 등식

SS_T의 자유도 = SS_{Tr}의 자유도 + SS_E의 자유도

$$(n\text{-}1) = (k\text{-}1) + (n\text{-}k)$$

총제곱합은 처리제곱합과 오차제곱합의 합과 같다는 것을 증명

총제곱합(SS_T)=처리제곱합(SS_{Tr}) + 오차제곱합(SS_E)

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(\overline{Y_i}-\bar{Y})+(Y_{ij}-\overline{Y_i})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{k=1}^{n_i}\left[(\overline{Y_i}-\bar{Y})^2+2(\overline{Y_i}-\bar{Y})(Y_{ij}-\overline{Y_i})+(Y_{ij}-\overline{Y_i})^2\right]^2$$

$$= \sum_{i=1}^{k}\left[n_i(\overline{Y_i}-\bar{Y})^2+2(\overline{Y_i}-\bar{Y})\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_i})+\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_i})^2\right]$$

0 $\because \sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_i}) = \sum_{j=1}^{n_i}(Y_{ij}-n_i\overline{Y_i}) = 0$

$$= \sum_{i=1}^{k}\left[n_i(\overline{Y_i}-\bar{Y})^2+\sum_{j=1}^{n}(Y_{ij}-\overline{Y_i})^2\right]$$

$$= \sum_{i=1}^{k}\sum_{k=1}^{n_i}(\overline{Y_i}-\bar{Y})^2+\sum_{i=1}^{k}\sum_{k=1}^{n_i}(Y_{ij}-\overline{Y_i})^2$$

집단간분산과 집단내분산의 비로 새로운 확률변수 생성

집단들의 모평균이 같다는 귀무가설로 새로운 확률변수(F)의 검정통계량(F_0)을 구함.

새로운 확률변수 $F = \frac{MS_{Tr}}{MS_E}$

요인 (factor)	편차제곱 합 (squared sum)	자유도 (degrees of freedom)	편차제곱 평균 (mean squared)	F검정통계량 (F statistic)
처리 (Between)	SS_{Tr}	$k-1$	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$ **집단간 분산**	$F_0 = \frac{MS_{Tr}}{MS_E}$ **집단들의 모평균이 같다는 귀무가설(0가설)**
오차 (Within)	SS_E	$n-k$	$MS_E = \frac{SS_E}{n-k}$ **집단내 분산**	
전체 (Total)	SS_T	$n-1$	$MS_T = \frac{SS_T}{n-1}$	

편차제곱합

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{..}})^2 \qquad SS_{Tr} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i.}} - \overline{Y_{..}})^2 = \sum_{i=1}^{k} n_i(\overline{Y_{i.}} - \overline{Y_{..}})^2 \qquad SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i.}})^2$$

F검정통계량

F검정통계량

$$F_0 = \frac{MS_{Tr}}{MS_E} = \frac{SS_{Tr}/(k-1)}{SS_E/(n-k)}$$

MS_{Tr} : 처리제곱평균, mean of squared treatment
MS_E : 오차제곱평균, mean of squared error
SS_{Tr} : 처리제곱합, sum of squared treatment
SS_E : 오차제곱합, sum of squared error
k : 집단의 개수
n : 전체집단의 표본의 크기

- MS_{Tr}을 MS_E로 나눈 비율이 크면 범주형 독립변수로 구분된 집단간 평균의 차이가 크다는 것을 의미합니다.
- 집단의 개수가 적고 각 집단의 표본크기가 크면 오차제곱합(SS_E)은 처리제곱합(SS_{Tr})에 비해 커집니다. 반대로 집단의 개수가 많고 각 집단의 표본크기가 작으면 오차제곱합(SS_E)은 처리제곱합(SS_{Tr})에 비해 작아집니다. 따라서 처리제곱합과 오차제곱합을 해당 자유도로 나누어 구한 오차제곱평균과 처리제곱평균을 비교합니다
- 검정통계량, F_0는 0을 기준으로 하는 처리제곱합을 사용함을 의미합니다.
- 오차제곱평균과 처리제곱평균은 각각의 자유도를 모수로 하는 카이제곱분포를 나타냅니다.
- 오차제곱평균과 처리제곱평균의 비는 두 자유도를 모수로 하는 F분포를 나타냅니다.

집단과 표본분산의 표집

- 집단은 확률변수값을 원소로 하는 집합
- 표집(Sampling distribution)은 집단에서 일정한 크기로 뽑을 수 있는 표본을 모두 뽑았을 때, 표본통계량을 원소로 하는 집합
  (ex. 표본평균 표집, 표본분산 표집)
- 표본분산 표집은 표본분산을 원소로 하는 집합
- 표본분산의 표집은 표본분산이 나타내는 확률분포로 표본크기가 커지면 0에 치우친 분포에서 중심극한정리에 의하여 종모양의 확률분포를 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

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

표본분산 표집

확률변수

$$S_X^2$$

표집크기

$$\infty$$

표본분산 표집

$$S_{X1}^2, S_{X2}^2, \cdots, S_{X\infty}^2$$

확률변수변환 : χ_{df}^2 분포

$$S_X^2 \to \chi_{df}^2$$

$$(n-1)\frac{S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2$$

여기서, n은 표본크기

표본분산 기대값(표집의 모평균)

$$\mathrm{E}[S_X^2] = \mu_{S_X^2} \sim \sigma_X^2$$

표본분산 표집의 모분산

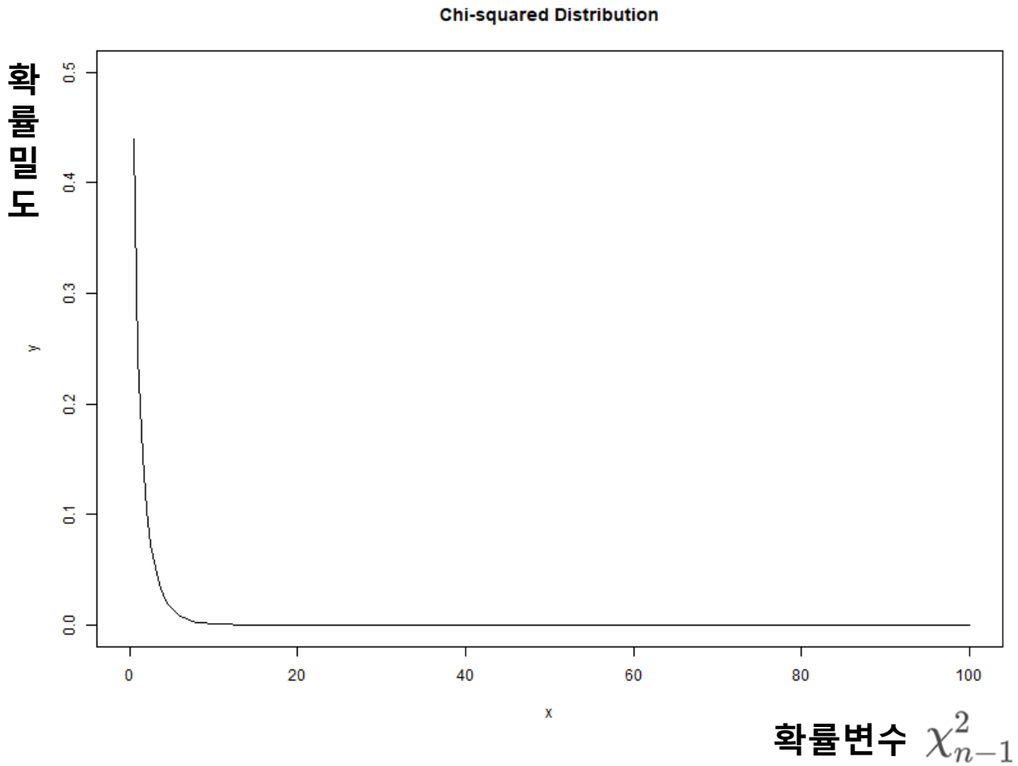
$$\mathrm{Var}(S_X^2) = \sigma_{S_X^2}^2 \sim \frac{2\sigma_X^4}{n-1}$$

여기서, n은 표본크기

표본분산 표집의 모표준편차

$$\mathrm{SD}(S_X^2) = \sigma_{S_X^2} \sim \sqrt{\frac{2\sigma_X^4}{n-1}}$$

표본분산을 무차원 확률변수인 카이제곱으로 변환



자유도를 1에서 100까지 증가시키면서 카이제곱분포의 확률밀도함수 관찰

- 확률변수가 X 이고 표본의 크기가 n 인 표본 :

$$X_1, X_2, \ldots, X_n$$

- 표본평균의 추정량(Estimator) :

$$\bar{X} = \frac{X_1 + X_2 +, \ldots, + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- 표본분산의 추정량(Estimator) :

$$S_X^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 +, \ldots, + (X_n - \bar{X})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- 표본분산의 기대값 :

$$\mathrm{E}[S^2] = \sigma^2$$

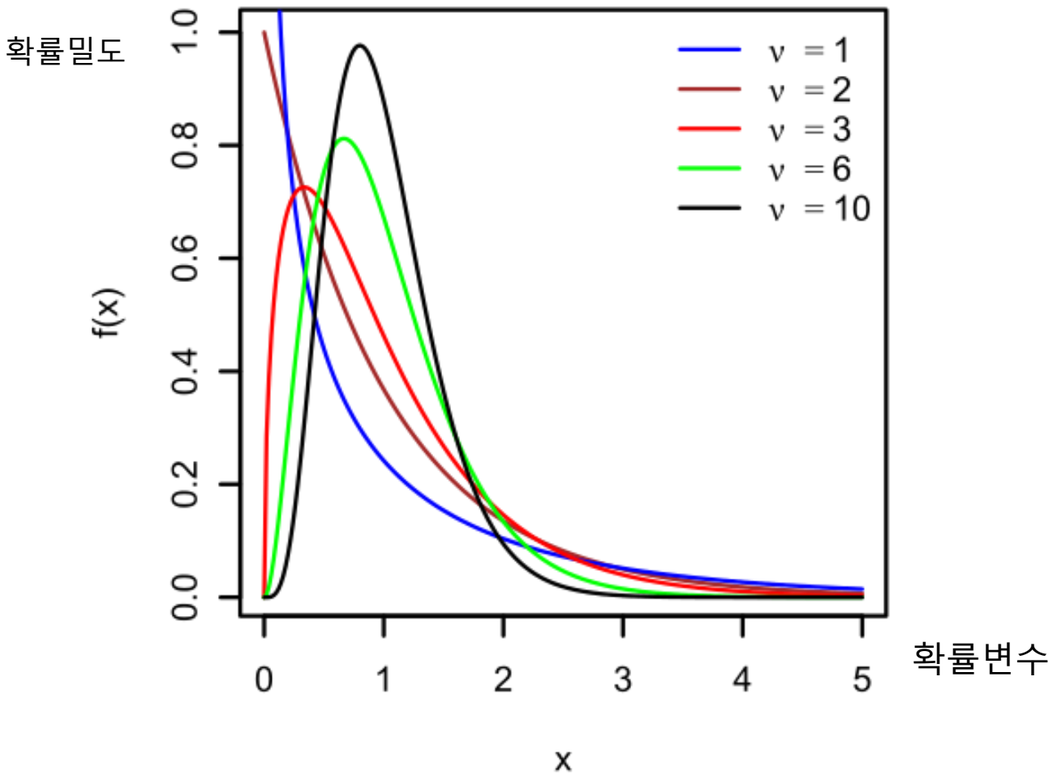
- 표본분산을 무차원 확률변수인 카이제곱으로 변환 :

$$\chi^2_{n-1} = (n-1)\frac{S^2}{\sigma^2}$$

- 카이제곱분포의 모수(parameter)인 자유도 :

$$df = n - 1$$

표본분산의 표준오차 -> 표본분산 표집의 모표준편차



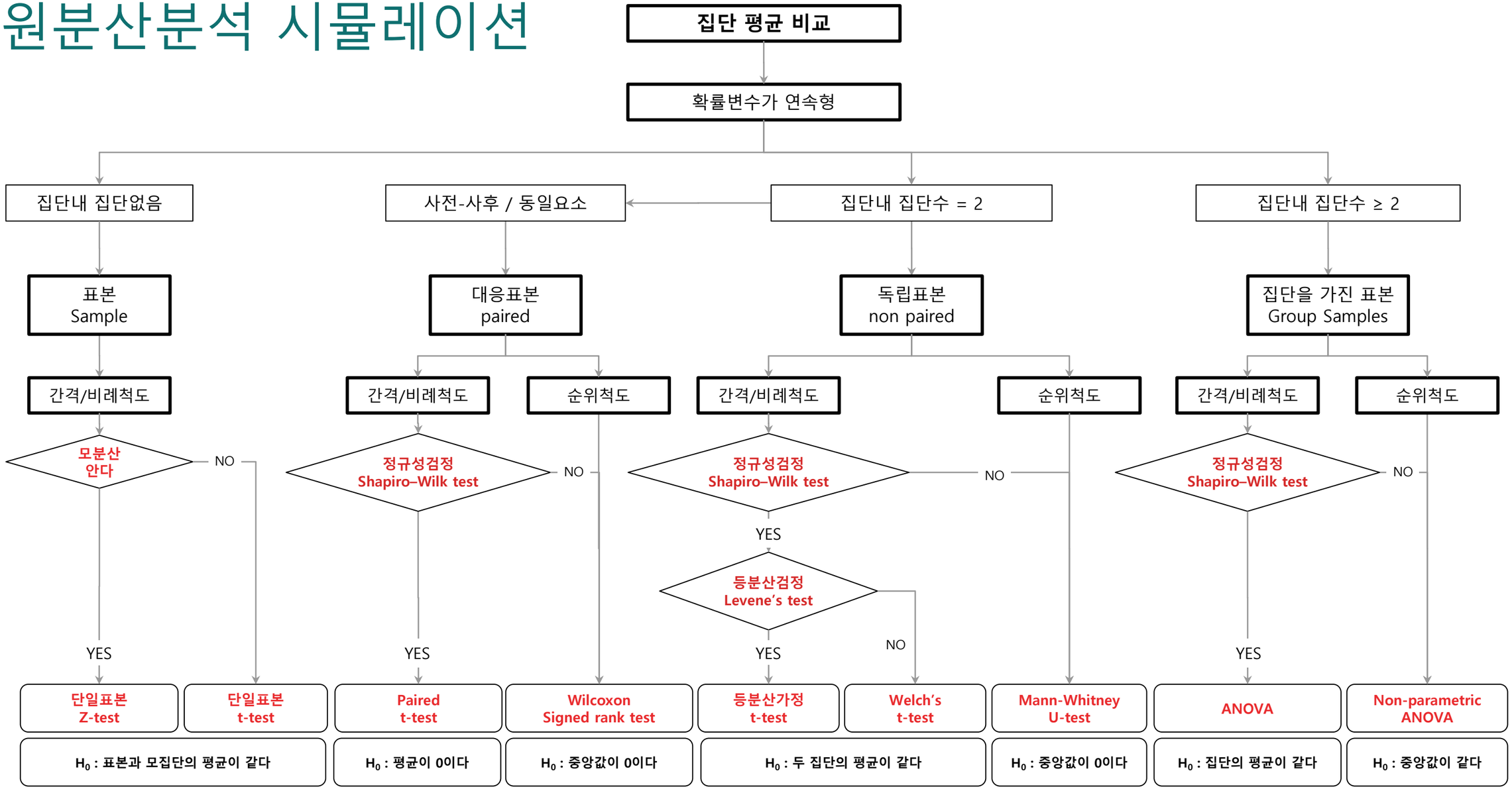
표본분산을 모분산으로 나눈값의 표본의 자유도에 따른 분포

표본분산 표집의 모표준편차 : $\mathrm{SD}(S_X^2) = \sigma_{S_X^2} \sim \sqrt{\frac{2\sigma_X^4}{n-1}}$

표본분산의 표준오차 : $SEV = \sigma_{S_X^2} = \sqrt{\frac{2\sigma_X^4}{n-1}} \sim \sqrt{\frac{2S_X^4}{n-1}}$

$\mathrm{SE}(S_X^2) = \sigma_{S_X^2} = \sqrt{\frac{2\sigma_X^4}{n-1}} \sim \sqrt{\frac{2S_X^4}{n-1}}$

일원분산분석 시뮬레이션



분산분석은 집단의 모평균의 동일성을 검정

분산분석(analysis of variance, ANOVA, AOV)은 2개 이상 집단의 모평균의 동일성(동질성)을 검정하는 방법

분석방법

1. 독립변수를 몇 개의 수준으로 나누고 각 수준에 따라 나누어진 집단 간의 평균이 동일한지를 검정한다.
2. 가설을 검정하기 위해 분산을 독립변수(또는 요인(factor))의 수준 차이에 기인한 부분과 우연적 오차에 의한 부분으로 분할한 다음, 전자가 후자보다 충분히 클 때 요인의 수준에 따라 집단 간 차이가 있는 것으로 판단한다.
3. 분산분석은 세 집단 이상을 비교하는 방법으로 두 집단 비교를 확산시킨 것이다.

기본가정

1. 정규성 가정 : 각 집단의 분포가 정규분포이다.
2. 등분산 가정 : 각 집단의 모분산이 같다.
3. 독립성 가정 : 집단에서 추출된 표본은 각각 독립이다.

일원분산분석은 독립변수가 1개

일원분산분석은 독립변수(원인, 요인변수)가 하나이고 종속변수(결과, 반응변수)가 1개인 분산분석

집단의 통계모델

$$Y_{ij} = \mu_Y + \epsilon_{ij} = \mu_Y + (\mu_{Y_i} - \mu_Y) + (Y_{ij} - \mu_{Y_i}) = \mu_Y + \alpha_i + \varepsilon_{ij}$$

여기서, Y_{ij}는 i번째 집단의 j번째 값

μ_Y는 전체집단의 모평균

ϵ_{ij}는 i번째 집단의 j번째 값(Y_{ij})과 전체집단의 모평균(μ_Y)과의 오차

μ_{Y_i}는 i번째 집단의 모평균

α_i는 i번째 집단의 모평균(μ_{Y_i})과 전체집단의 모평균(μ_Y)간의 편차

ε_{ij}는 i번째 집단의 j번째 값(Y_{ij})과 i번째 집단의 모평균(μ_{Y_i})과의 오차

k가 집단의 개수라면, $i=1,2,\cdots,k$

N_i가 i번째 집단의 크기라면, $j = 1, 2, \cdots, N_i$

일원분산분석의 독립변수는 집단을 구분하는 범주형 변수

원인변수값에 의해 구분된 각 집단의 변동을 표본을 통해 관측할 수 있습니다.

표본의 통계모델

$$Y_{ij} = \bar{Y} + (\overline{Y_i} - \bar{Y}) + (Y_{ij} - \overline{Y_i})$$

여기서, Y_{ij}는 i번째 집단의 j번째 관측값

$\bar{Y}$는 전체집단의 표본평균

$\overline{Y_i}$는 i번째 집단의 표본평균

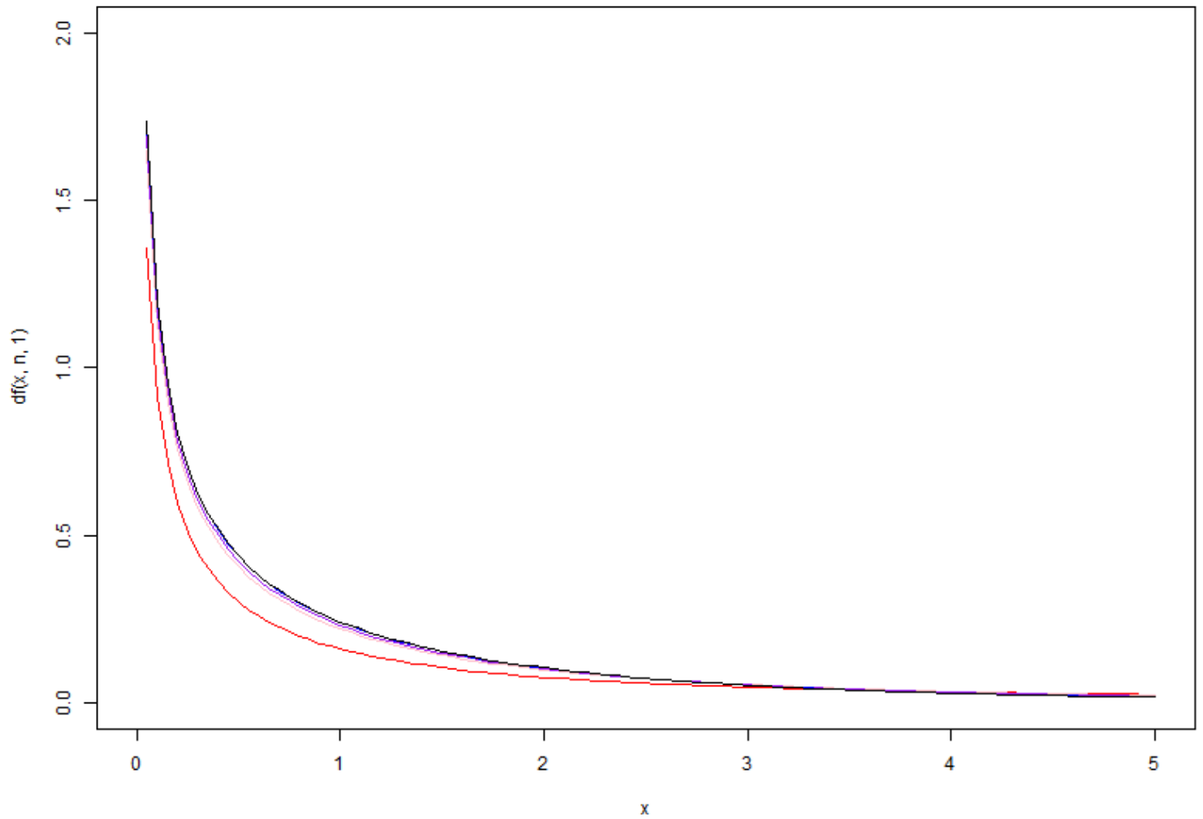
$(\overline{Y_i} - \bar{Y})$는 i번째 집단의 표본평균$(\overline{Y_i})$과 전체집단의 표본평균$(\bar{Y})$간의 편차

$(Y_{ij} - \overline{Y_i})$는 i번째 집단의 j번째 관측값(Y_{ij})과 i번째 집단의 표본평균$\overline{Y_i}$과의 오차 : 잔차$(X_{residual}, X_r)$

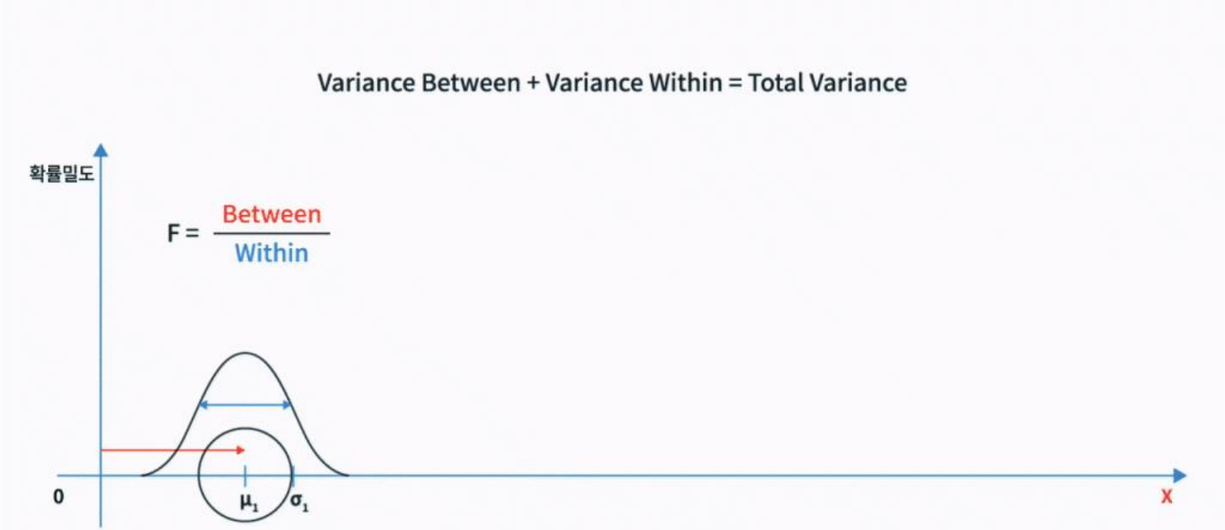
k는 집단의 개수라면, $i=1,2,\cdots,k$

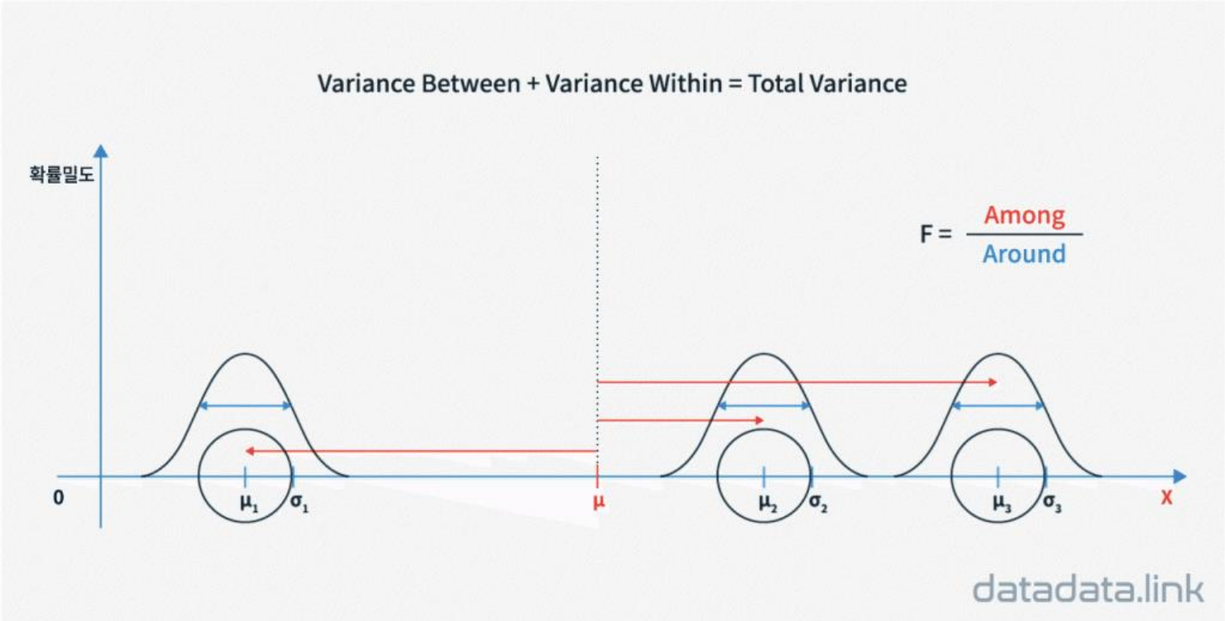
n_i가 i번째 집단의 표본크기라면, $j = 1, 2, \cdots, n_i$

확률변수 F로 변환

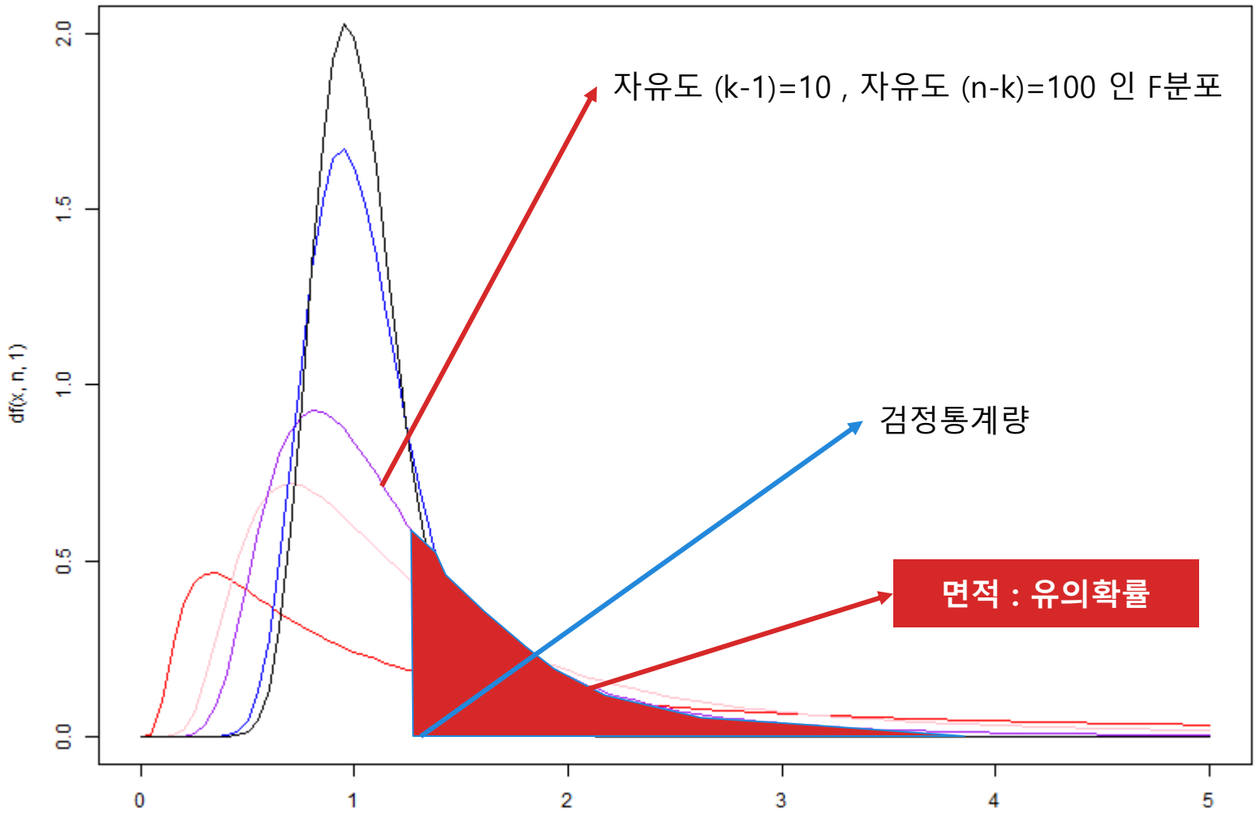


d_2 가 1, 5, 10, 50, 100 일때 각각 d_1 을 1에서 100으로 증가시킬 때 F분포의 변화





표본데이터로 F분포에서의 검정통계량과 유의확률 구하기



d_2 가 10이고 d_1 가 100 인 F분포에서의 유의확률

검정통계량

$$F_0 = \frac{MS_{Tr}}{MS_E} = \frac{SS_{Tr}/(k-1)}{SS_E/(n-k)}$$

여기서, k는 전체집단을 이루는 집단의 수, n은 전체집단의 표본크기

제곱합

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{..}})^2$$

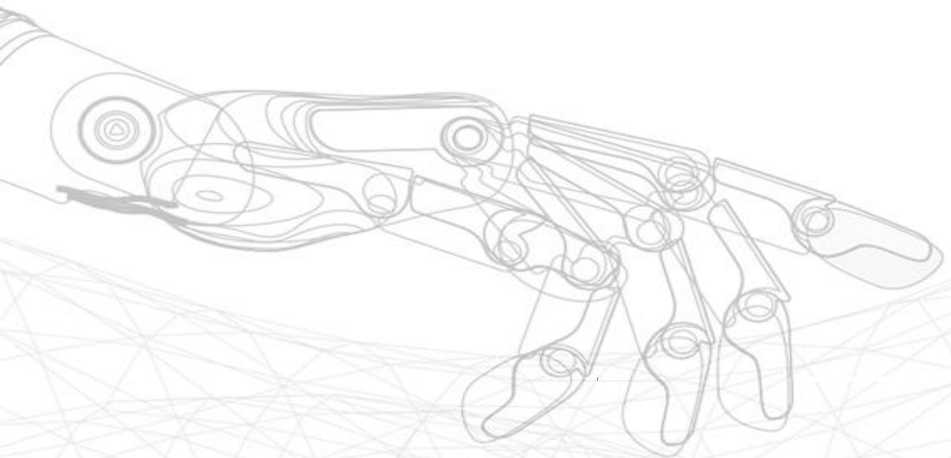
$$SS_{Tr} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y_{i.}} - \overline{Y_{..}})^2 = \sum_{i=1}^{k} n_i(\overline{Y_{i.}} - \overline{Y_{..}})^2$$

$$SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_{i.}})^2$$

기각기준

$F_0 > F_{k-1,n-k;\alpha}$ 이면 H_0 를 기각

여기서, F분포의 모수인 분자, 분모의 자유도는 (k-1), (n-k)



감사합니다

www.datadata.link