# 한우 도체중의 ‘일원분산분석 F검정’과 ‘독립표본 t검정’

Seung Hwan Lee (https://orcid.org/0000-0003-1508-4887)

Chungnam National University, Daejeon, Republic of Korea

Corresponding author: Seung Hwan Lee (slee46@cnu.ac.kr, genomicselection46@gmail.com)

### 초록(Abstract)

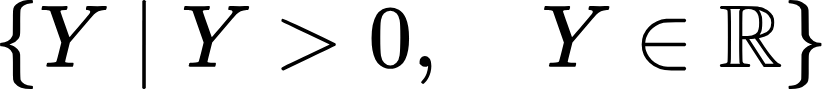
한우의 도체중은 연속형 확률변수로 정규분포를 따릅니다. 따라서 모집단의 도체중은 정규분포를 나타냅니다. 모집단내의 집단 수를 k로 하고 각 집단의 분산은 동일하다고 가정합니다. 표본은 한우 n마리를 무작위로 추출하여 모집단을 대표합니다. 자유도는 표본내 개체와 집단에 적용되어, 표본내 집단의 자유도는 (k-1), 표본내 개체의 자유도는 (n-k)가 됩니다. 변동은 집단의 변동(신호)과 집단내 개체의 변동(노이즈)으로 나눌 수 있으며, 변동과 자유도로 집단의 분산(집단간분산)과 개체의 분산(집단내분산)을 계산합니다. 이 집단간분산과 집단내분산의 비는 F분포를 따릅니다. 독립된 두 집단의 평균차이를 검정할 때, t분포를 따르는 t통계량의 제곱은 F분포를 따릅니다. t분포의 모수인 자유도와 F분포의 분모의 모수인 자유도는 같고 F분포의 분자의 자유도는 1입니다.

### 주제어(Keywords)

개체, 집단, 모집단, 표본, 정규분포, 카이제곱분포, F분포, t분포

### 모델링: 확률모델

#### 개체(individual)

* 개체: 한우
  + 한우의 도체중 -> 한우의 연속형 속성 -> 결과변수 -> 확률변수
    - 한우의 도체중 = 
      * 한우 도체중은 확률변수: 독립동일분포 가정 (Independent Identificate Distribution)
        + 독립동일분포의 확률분포형태: 정규성 가정 -> 정규분포

한우 도체중의 기대값: 평균 (정규분포함수의 매개변수)

한우 도체중의 분포값의 기대값: 분산 (정규분포함수의 매개변수)

#### 집단(group, category)

* 집단: 지역
  + 한우의 산지 -> 한우의 범주형 속성 -> 원인변수
    - 한우의 산지 = {A, B, C}
      * 집단의 수: k=3
        + 개체는 집단의 한 곳에만 속함: 집단의 독립성 가정
        + 각 집단에서 도체중의 분산은 동일: 등분산 가정

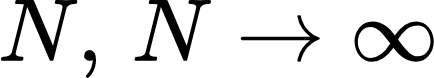
각 집단에서 도체중의 분포: 정규분포

* 집단: 출생년도
  + 한우의 출생년도 -> 한우의 범주형 속성 -> 원인변수
    - 한우의 출생년도 = {2011년, 2012년}
      * 집단의 수: k=2
        + 개체는 집단의 한 곳에만 속함: 집단의 독립성 가정
        + 각 집단에서 도체중의 분산은 동일: 등분산 가정

각 집단에서 도체중의 분포: 정규분포

### 모델링: 통계모델

#### 모집단(population)

* 모집단(population): A, B, C지역의 모든 한우
  + 모집단의 도체중 분포: 정규분포(충분히 큰 모집단크기)
  + 모집단크기: 
  + 모집단내 각 집단에서 도체중의 분포: 정규분포
  + 모집단내 각 집단의 도체중의 분산은 동일: 등분산 가정
  + 개체는 모집단내 집단의 한 곳에만 속함: 집단의 독립성 가정
  + 모집단내 집단의 수: k
  + 모집단내 집단: A, B, C 지역, k=3
  + 모집단내 집단의 상대빈도(확률): PA, PB, PC
  + PA+PB+PC=1
  + 모집단내 집단: 2011년, 2912년 출생년도, k=2
  + 모집단내 집단의 상대빈도(확률): P2011, P2012
  + P2011+P2012=1

#### 표본

* 표본: 모집단에서 무작위(랜덤)로 추출한 한우
  + 표본의 도체중 분포: 정규분포 (표본추출의 무작위, 충분히 큰 표본크기)
  + 표본크기: n
  + 표본내 각 집단에서 도체중의 분포: 정규분포 (표본추출의 무작위, 충분히 큰 표본크기)
  + 표본내 각 집단의 도체중의 분산은 동일: 등분산 가정
  + 개체는 표본내 집단의 한 곳에만 속함: 집단의 독립성 가정
  + 표본내 집단의 수: k=3
    - 표본내 집단: A, B, C지역
      * 표본내 집단의 크기: nA, nB, nC
      * nA+nB+nC=n
    - 표본내 집단: 2011년, 2012년 출생년도, k=2
      * 표본내 집단의 크기: n2011, n2012
      * n2011+n2012=n

#### 일변량 표본(univariate sample)의 자유도(degree of freedom) - 집단이 있는 경우

* 개체가 i번째 집단으로 묶이면 i번째 집단이 생성되고 i번째 집단의 자유도는 1 -> i번째 집단내 개체의 자유도는 1이 감소하여 (ni-1)
* 표본내 개체의 자유도는 집단내 개체의 자유도의 합 -> 표본내 개체의 자유도는 (n-k)
* 집단이 표본집단(전체집단)으로 묶이면 표본집단이 생성되고 표본집단의 자유도는 1 -> 표본내 집단의 자유도는 집단수에서 1이 감소하여 (k-1)이고, 또한, 표본의 자유도는 표본크기에서 1이 감소하여 (n-1) -> 표본의 자유도는 표본내 집단의 자유도와 표본내 개체의 자유도의 합, (n-1)=(k-1)-(n-k)

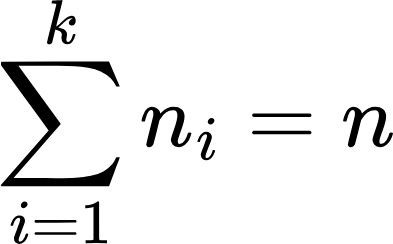
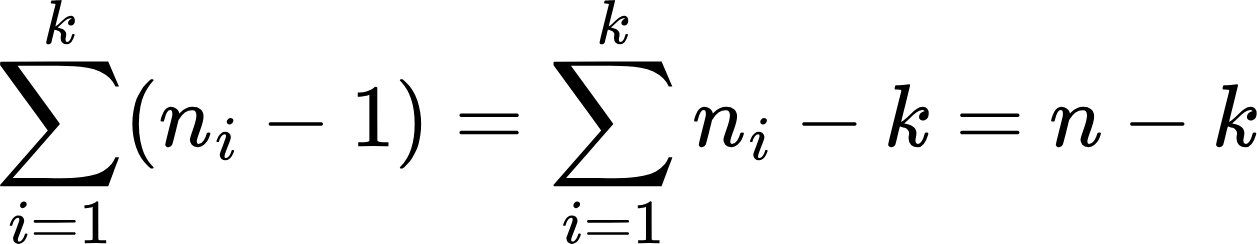
#### 단일표본(집단이 1개)에서의 자유도

* 표본내 개체의 수 (표본크기) = n
* 표본내 개체의 자유도 (표본의 자유도) = n-1

#### 대응표본(대응된 두 집단)에서의 자유도

* 표본내 개체의 수 (표본크기) = n
* 표본내 집단의 자유도 = 0: 귀무가설에서 집단은 1개
* 표본내 개체의 자유도 (표본의 자유도) = n-1

#### 독립표본(독립된 집단이 2개 이상)에서의 자유도

* 표본내 개체의 수(표본크기) = n
  + 표본내 개체의 자유도 (표본의 자유도) = n-1
* 표본내 집단의 수 = k
  + 표본내 집단의 자유도 (집단간 자유도) = k-1
* 표본내 집단 i의 개체의 수 = ni : 
  + 집단 i내 개체의 자유도 (집단 i의 자유도) = ni-1
* 집단내 개체의 자유도의 합 (집단내 자유도) = 
* 표본의 자유도 = 집단간 자유도 + 집단내 자유도 = (k-1)+(n-k)=n-1

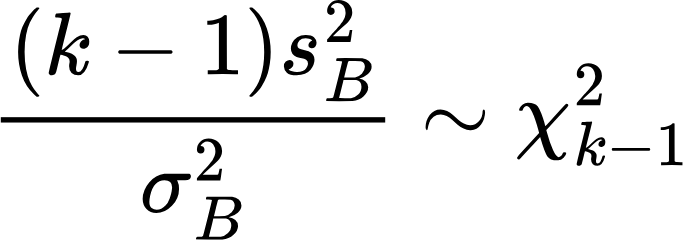
### 모델링: 새확률변수

#### 표본변동(sample variation)의 분할

* 표본내 집단의 변동과 표본내 개체의 변동으로 분할: 집단간변동과 집단내변동으로 분할
* 범주형 원인변수에 의해 집단간변동이 나타나고 연속형 결과변수의 독립동일분포가정에 의해 집단내변동이 나타난다고 모델링

#### 변동의 표준화: 카이제곱

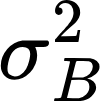
확률변수인 집단간분산을 집단간분산의 모분산으로 표준화하면 자유도가 (집단수-1)인 카이제곱분포를 따르는 확률변수가 된다.



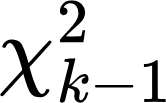
여기서, k는 집단수

k-1은 표본내 집단의 자유도

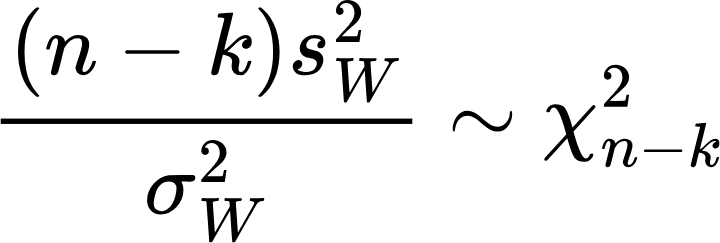
SB2은 집단간분산

은 집단간분산의 모분산

(k-1)SB2은 표본내 집단의 변동(집단간변동)

은 (k-1)인 자유도를 매개변수로 하는 카이제곱분포

확률변수인 집단내변동을 집단내분산의 모분산으로 표준화하면 자유도가 (표본크기-집단수)인 카이제곱분포를 따르는 확률변수가 된다.

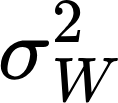


여기서, n은 표본크기

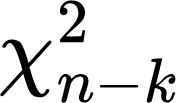
k는 집단수

n-k는 표본내 개체의 자유도

sW2은 집단내분산

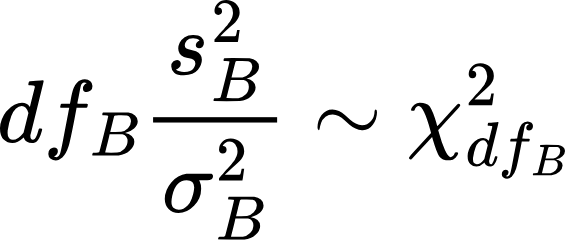
은 집단내분산의 모분산

(n-k)sW2은 표본내 개체의 변동(집단내변동)

은 (n-k)인 자유도를 매개변수로 하는 카이제곱분포

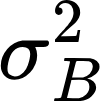
#### 집단내분산에 대한 집단간분산의 비: F

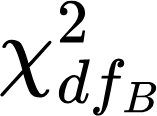
표준화된 집단간변동은 확률변수이며 카이제곱분포를 따른다.



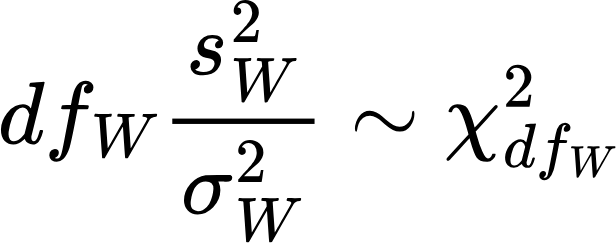
여기서, dfb는 표본내 집단의 자유도

sB2은 집단간분산: 표본의 집단내분산

은 모집단간분산: 모집단의 집단간분산

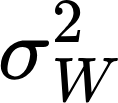
는 dfB를 매개변수로 하는 카이제곱분포

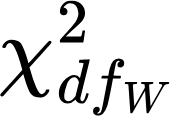
표준화된 집단내변동은 확률변수이며 카이제곱분포를 따른다.



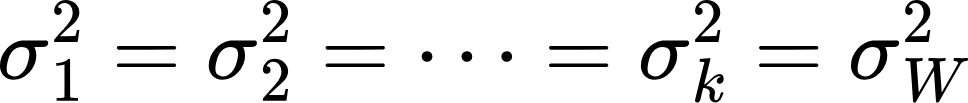
여기서, dfW는 표본내 개체의 자유도

sW2은 집단내분산: 표본의 집단내분산

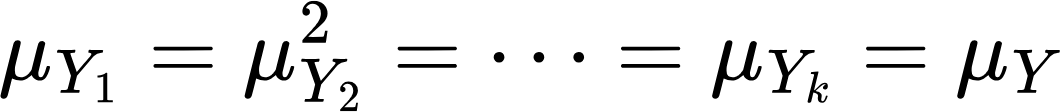
은 모집단내분산: 모집단의 집단내분산

는 dfW를 매개변수로 하는 카이제곱분포

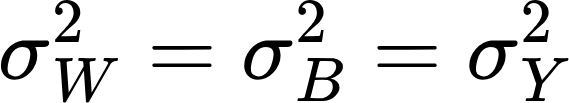
등분산가정 : 모집단의 각 집단(1집단, 2집단, … , k집단 )의 분산이 같다.



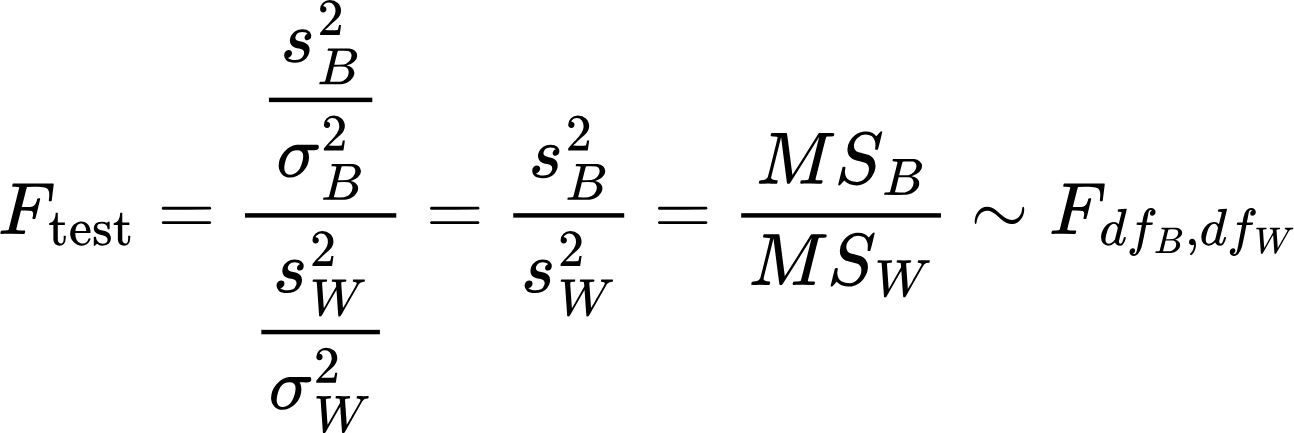
귀무가설 : 모집단의 각 집단(1집단, 2집단, … , k집단 )의 평균이 같다. -> 모집단내 집단(group)의 평균과 모집단의 전체평균이 같다.



모집단의 각 집단(group)의 평균이 같다. -> 모집단의 집단내분산은 집단간분산과 같다. -> 모집단의 집단내분산은 모집단의 분산과 같다.



표본의 집단간분산/집단내분산의 F검정통계량은 새확률변수이며 F분포를 나타낸다.



확률변수가 정규분포인 경우 F분포를 나타내는 새확률변수

| 표본내 집단의 크기 | 변동 | 자유도 | 분산 | 새확률변수 |
| --- | --- | --- | --- | --- |
| {"aid":null,"id":"23","code":"$$n=\\sum\\limits_{i=1}^{k}n_{i}$$","font":{"size":9,"color":"#000000","family":"Malgun Gothic"},"backgroundColorModified":false,"backgroundColor":"#ffffff","type":"$$","ts":1726183175785,"cs":"kj4t3z65v5AoNn6HTCesug==","size":{"width":52,"height":32}}  여기서, n은 표본내 개체수  k는 표본내 집단수  ni는 i번째 집단의 크기 | 표본내 집단의 변동 SSB  표본내 개체의 변동 SSW | 표본내 집단의 자유도 dfB=k-1  표본내 개체의 자유도 dfW=n-k | 집단간분산  {"code":"$$MS_{B}=\\dfrac{SS_{B}}{df_{B}}$$","backgroundColor":"#ffffff","aid":null,"backgroundColorModified":false,"font":{"size":9,"family":"Malgun Gothic","color":"#000000"},"type":"$$","id":"24","ts":1726183262972,"cs":"Sh3t0GmzeqdG5jxS9gYgKA==","size":{"width":69,"height":26}}  집단내분산  {"backgroundColorModified":false,"font":{"color":"#000000","size":9,"family":"Malgun Gothic"},"backgroundColor":"#ffffff","type":"$$","id":"26","code":"$$MS_{W}=\\dfrac{SS_{W}}{df_{W}}$$","aid":null,"ts":1726183314412,"cs":"b5eXpA5O7TCbNY5rVO9CLA==","size":{"width":74,"height":26}} | 집단내분산에 대한 집단간분산의 비  {"backgroundColor":"#ffffff","id":"27","code":"$$\\dfrac{MS_{B}}{MS_{W}}$$","backgroundColorModified":false,"type":"$$","aid":null,"font":{"size":9,"family":"Malgun Gothic","color":"#000000"},"ts":1726183337346,"cs":"9Loz3k6vOvopOawJDS8r9A==","size":{"width":32,"height":26}} |

### 데이터

#### 한우의 연속형 속성과 범주형 속성 그리고 원인변수와 결과변수 그리고 확률변수

* 개체: 한우
  + 개체의 연속형 속성(결과변수): 한우의 도체중
    - 결과변수의 변동: 도체중의 변동
  + 개체의 범주형 속성(원인변수): 한우의 출생년도, 한우의 출생지
    - 결과변수 변동의 분할: 원인변수에 의한 집단의 변동(신호, 처리)과 개체의 변동(노이즈, 오차)
      * 표본내 집단의 변동(SSB): 표본평균으로부터 집단평균의 편차의 제곱의 합
      * 표본내 개체의 변동(SSW): 각 집단내 개체의 변동의 합
        + 각 집단내 개체의 변동: 각 집단의 평균으로부터 개체의 편차의 제곱의 합

### 데이터분석: 일원분산분석 F검정

일원분산분석표

| 변동 | 변동 표기  (Squared Sum) | 자유도 표기  (degrees of freedom) | 분산 표기  (Mean Squared) | F검정통계량  (F statistic) |
| --- | --- | --- | --- | --- |
| 표본내 집단의 변동  (Between variation) | 집단간변동 SSB | 집단의 자유도  dfB=k-1 | 집단간분산  {"id":"29","code":"$$MS_{B}=\\dfrac{SS_{B}}{k-1}$$","backgroundColor":"#ffffff","aid":null,"backgroundColorModified":false,"font":{"size":9,"color":"#000000","family":"Malgun Gothic"},"type":"$$","ts":1726183954558,"cs":"OMQsYpvO3AxdKKmNMunm/Q==","size":{"width":74,"height":24}} | {"font":{"size":12,"family":"Arial","color":"#000000"},"code":"$$F_{0}=\\dfrac{MS_{B}}{MS_{W}}$$","id":"28","backgroundColor":"#ffffff","backgroundColorModified":false,"type":"$$","aid":null,"ts":1726183914828,"cs":"RfMoq1netOgau/6BtyoqVQ==","size":{"width":96,"height":41}} |
| 표본내 개체의 변동  (Within variation) | 집단내변동 SSW | 개체의 자유도  dfw=n-k | 집단내분산  {"code":"$$MS_{W}=\\dfrac{SS_{W}}{n-k}$$","aid":null,"id":"30","backgroundColorModified":false,"type":"$$","font":{"family":"Malgun Gothic","color":"#000000","size":9},"backgroundColor":"#ffffff","ts":1726184048656,"cs":"Jovb44aEEtVNCkSjFgvREw==","size":{"width":77,"height":24}} |
| 표본의 변동  (Total variation)  표본내 총변동 = 표본내 집단의 변동 + 표본내 개체의 변동 | 총변동 SST  SST=SSB+SSW | 총자유도  dfT=n-1  dfT=dfB+dfW | 표본분산  {"aid":null,"id":"31","type":"$$","font":{"family":"Malgun Gothic","size":9,"color":"#000000"},"backgroundColor":"#ffffff","backgroundColorModified":false,"code":"$$MS_{T}=\\dfrac{SS_{T}}{n-1}$$","ts":1726184069603,"cs":"rlXJQedBC+qWux1nRTMe2A==","size":{"width":74,"height":24}}  {"aid":null,"type":"$$","code":"$${df_{T}}\\cdot{MS_{T}}={df_{B}}\\cdot{MS_{B}}+{df_{W}}\\cdot{MS_{W}}$$","id":"32","font":{"size":9,"family":"Malgun Gothic","color":"#000000"},"backgroundColorModified":false,"backgroundColor":"#ffffff","ts":1726184094389,"cs":"0gPgCGS6IpkssypMkjfEVQ==","size":{"width":202,"height":10}} |

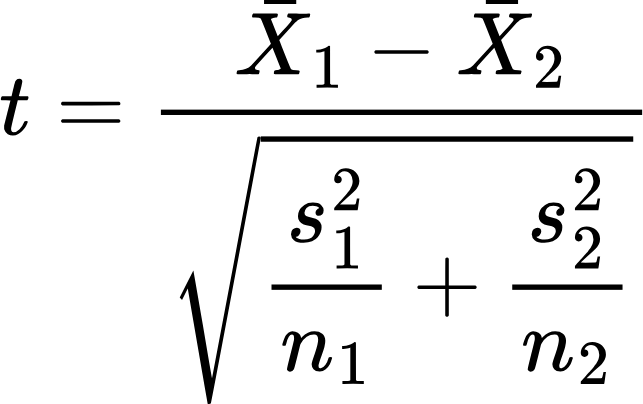
### 

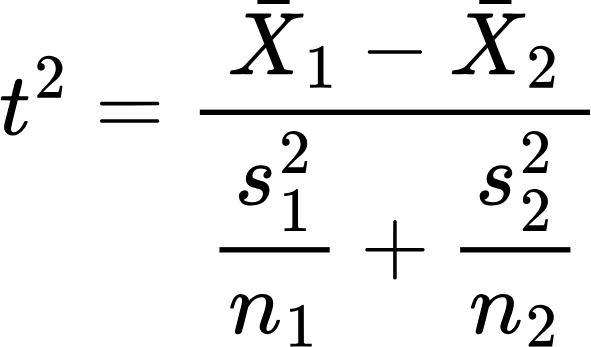
### t검정과 F검정의 관계

#### 독립표본 t검정과 일원분산분석 F검정

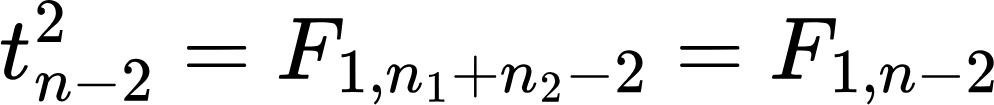
* 일원분산분석은 2개 이상의 독립된 집단이 있는 독립표본에서 행한다.
* 독립표본 t검정은 독립표본에서 독립된 집단이 2개 있는 경우이고 일원분산분석 F검정은 독립표본에서 독립된 집단이 2개 이상 있는 경우이다. 표본내에 독립된 집단이 2개 있는 경우에는 독립표본 t검정의 결과와 일원분산분석 F검정의 결과가 같다.
* 독립표본에서 독립된 두 집단의 평균의 차이를 분석하는 것은 일원분산분석 F검정에서 독립된 두 집단의 평균의 분산을 분석하는 것과 같다.

독립표본에서 독립된 두 집단의 평균차이를 검정할 때, 다음식이 성립한다.





독립표본에서 확률변수 t와 확률변수 F의 관계식은 다음과 같다. 증명은 “참고문헌 4”



여기서, n은 표본크기: n=n1+n2

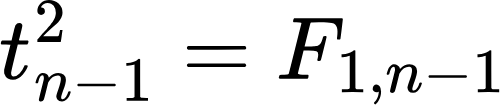
n1은 1번째 집단의 크기

n2는 2번째 집단의 크기

#### 대응표본의 t검정통계량과 F검정통계량의 관계 - 귀무가설

대응표본 t검정은 대응된 두 집단의 비교이다. 두 집단이 대응되어 있으므로 대응된 원소는 1개의 집단이라고 볼 수 있으며 그 집단의 원소수는 대응된 쌍의 수와 같게 된다. 따라서 표본내 집단의 자유도는 1이 되고 대응된 쌍의 자유도는 “쌍의 수 -1″이 된다.

대응표본에서 t통계량과 F통계량의 관계식은 다음과 같다.



여기서, n은 대응된 쌍의 수

#### 집단(group)이 2개 있는 독립표본의 t검정통계량과 F검정통계량의 관계 - 귀무가설

독립표본 t검정은 독립된 두 집단의 비교이다. 두 집단이 독립되어 있고 집단마다 평균이 있다. 표본내 집단의 자유도는 2-1=1 이고 개체의 자유도는 두 집단의 움직이는 평균 2개를 뺀 (n-2)가 된다.

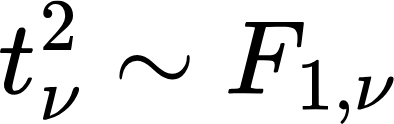
독립표본에서 확률변수 t와 확률변수 F의 관계식은 다음과 같다.

#### {"id":"38","aid":null,"font":{"color":"#000000","size":13,"family":"Malgun Gothic"},"backgroundColor":"#ffffff","code":"$$t_{n-2}^{2}=F_{1, n-2}$$","backgroundColorModified":false,"type":"$$","ts":1726211208746,"cs":"/BrnX/tRr/ISeFXQRRdR+A==","size":{"width":97,"height":20}}

여기서, n은 표본크기: 표본내 두 집단의 크기의 합 n=n1+n2

#### 확률변수 t와 확률변수 F의 일반화된 관계

확률변수 t와 확률변수 F분포의 관계를 일반화하면 다음과 같습니다



여기서, {"backgroundColorModified":false,"aid":null,"id":"40","font":{"family":"Malgun Gothic","size":11,"color":"#000000"},"code":"$$\\nu$$","type":"$$","backgroundColor":"#ffffff","ts":1726211277592,"cs":"uiaQmz8KcdxK3Dsy2tKWtA==","size":{"width":6,"height":5}}는 자유도

#### 집단평균에 대한 귀무가설에서 집단과 개체의 분산은 같은 것이 아닌 중첩

* 독립표본 t검정에서 두 집단의 모평균의 차이가 없다는 귀무가설은 두 집단의 모평균이 같다는 것이다. 이 경우에도, 두 집단평균의 분산은 나타나며 이 분산은 개체의 분산에서 기인한 것이라고 볼 수 있다. 즉, 집단평균의 귀무가설 하에서 개체의 분산과 두 집단평균의 분산은 같은 것이 아닌 중첩되어 있다고 볼 수 있다.
* 일원분산분석 F검정에서 여러 집단의 모평균의 차이가 없다는 귀무가설은 여러 집단의 모평균이 같다는 것이다, 이 경우에도 여러 집단평균의 분산은 나타나며 이 분산은 개체의 분산에서 기인한 것이라고 볼 수 있다. 즉, 집단평균의 귀무가설 하에서 개체의 분산과 여러 집단평균의 분산은 같은 것이 아닌 중첩되어 있다고 볼 수 있다.
* 대응표본 t검정에서 대응된 두 집단의 차이가 주어진 값이라는 귀무가설은 두 집단의 대응된 개체의 차이는 주어진 값이라는 것이다. 이 경우에도 대응된 집단의 분산은 나타나며 이 분산은 개체의 분산에서 기인한 것이라고 볼 수 있다. 즉, 대응된 집단의 귀무가설 하에서 개체의 분산돠 대응된 집단평균의 분산은 같은 것이 아닌 중첩되어 있다고 볼 수 있다.

### 참고자료(References)

1. [A, B, C지역 한우의 도체중 비교](https://www.datadata.link/a2-3/)

2. [범주형 원인변수에 의한 집단의 변동](https://www.datadata.link/s-13-3-1/)

3. [한 범주형 범주로 구분된 여러 집단 모평균 비교: 일원분산분석 F검정](https://www.datadata.link/s2/)

4. [독립된 두 집단의 모평균 비교: 독립표본 t검정](https://www.datadata.link/s-23-1-2/)