사회통계
일원분산분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 분산분석의 기초
2. 일원분산분석
3. 일원분산분석의 활용

분산분석의 기초

분산분석의 기초

Fisher는 F 분포를 활용하는 분산분석도 만들었다.

- Fisher는 F 분포를 만들었을 뿐만 아니라, 그 원리를 절묘하게 활용하는 분산분석(Analysis of Variance; ANOVA)도 개발했다.
- Fisher는 젊은 시절에 대학 대신 농업연구소 Rothamsted에서 일하면서 체계적인 실험설계(experimental design)의 근본 원리로 이른바 피셔의 3원칙(Fisher's Three Principles)인 반복(replication), 무작위화(randomization), 국소화(blocking)를 제시한다.
- 이 원칙들이 오늘날에는 당연하게 여겨지지만 당시에는 상당한 반감을 샀다.

분산분석의 기초

분산분석으로 어떤 변수의 분산에 관한 가설검정을 수행한다.

- 이전에 설명한 분산의 추정과 비교는 사실 분산분석을 위한 준비에 불과하다!
- 분산분석은 실험설계에서 가장 많이 사용되는 통계 분석방법론 중 하나이다. 특히 심리학, 보건의료 등에서는 굉장히 심도있게 연구된다(Why?).
- 분산분석은 나름 독특한 접근방법, 용어, 기법들의 집합이므로 처음에는 낯설게 느껴질 수 있다.

분산분석의 기초

- 일원분산분석(one-way ANOVA)은 분산분석에 속한 여러 기법들의 가장 기초가 된다.
- 분산분석의 기본적인 개념을 학습하는데 유용할 뿐 아니라, 실제로도 여러가지 맥락에서 자주 사용된다.
- 사실 심리학을 제외한 다른 사회과학 분야에서는 실험자료(experimental data) 대신 관찰자료(observational data)가 더 폭넓게 사용된다. 다시 말해, 분산분석이 폭넓게 사용되기 어렵다.
- 그러므로 우리는 오로지 (분산분석의 기초인) 일원분산분석만 배운다.

분산분석의 기초

일원분산분석의 데이터 구조가 어떻게 생겼는지 기억하자.

- 완전무작위설계(completely randomized design)에 따라 27명의 실험 참가자(subjects)에 대해 3가지의 서로 다른 실험처리(treatment)를 "같은 수 만큼 반복" 한다고 하자.

Treatment 1	Treatment 2	Treatment 3	
y_{11}	y_{12}	y_{13}	
y_{21}	y_{22}	y_{23}	
$\vdots$	$\vdots$	$\vdots$	
y_{91}	y_{92}	y_{93}	
$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\bar{y}_{\cdot 3}$	$\bar{y}_{\cdot\cdot}$

- 실험처리별 그룹이 3개라는 점, 각 그룹별로 9개(=27/3) 씩 표본이 들어있다는 점, 그리고 $\bar{y}_{\cdot j}$ 와 $\bar{y}_{\cdot\cdot}$ 같은 평균값에 주목하자.

분산분석의 기초

분산분석의 기초적인 용어에 먼저 친숙해질 필요가 있다.

- 분산분석의 맥락에서 독립변수는 요인(factor)이라고 불리운다. 분산분석에서 요인은 반드시 질적 변수이고 주로 실험처리 그룹의 식별자(identifiers)가 된다(Why?). 이 때문에 종종 그룹화 요인(grouping factor)이라고도 불리운다.
- 종속변수는 결과(outcome) 또는 반응 변수(response variable)라고 불리운다. 분산분석에서 이것은 반드시 양적 변수가 된다.
- 일원분산분석의 경우 분석에 사용되는 요인과 반응변수는 각각 하나씩이다.

분산분석의 기초

분산분석은 2개 이상의 모평균에 관한 가설검정에 사용된다.

- t 검정을 통해 두 모평균의 차이 $\mu_1 - \mu_2$에 관한 가설검정을 수행할 수 있었다.
- 분산분석을 통해서는 둘 이상의 집단에서 모든 모평균이 같은지 $\mu_1 = \mu_2 = ... = \mu_j$에 관한 가설검정도 수행할 수 있다.
- 분산분석은 t 검정보다 훨씬 일반화된 분석기법인 셈이다(Why?).
- 그러나 연구나 실무에서는 범주형 독립변수에서 주어진 범주의 수가 2개 일 때(예컨대 성별, 성인 여부, 대졸 여부 등)는 주로 t 검정을, 주어진 범주의 수가 3개 이상일 때(예컨대 최종학력별, 지역별, 고용조건별 등)만 분산분석을 사용한다.

일원분산분석

일원분산분석

일원분산분석은 비교적 단순한 논리 구조를 가지고 있다.

- 일원분산분석의 모집단에 관한 수리모형은 다음과 같다.

$$Y_{ij} = \mu_j + e_{ij}$$

- 이때 e_{ij}는 오차항(error term)이다. 오차는 무작위적(random)이라고 가정된다.
- 처리효과(treatment effect)는 $\alpha_j = \mu_j - \mu$이므로(Why?),

$$Y_{ij} = \mu + \alpha_j + e_{ij}$$

- 다시 말해, (종속변수가 되는) 모집단 개체의 관찰값 Y_{ij}은 전체 평균 μ와 처리효과 α_j, 그리고 오차항 e_{ij}으로 분해된다.
- α_j가 크고 e_{ij}가 작을 때 처리집단 간 의미있는 차이가 있다고 말할 수 있다(Why?).

일원분산분석

수학적으로 일반화되어 어려워 보이니 쉬운 예제를 통해 살펴보자.

- 가령 첫번째 열(column)에서 y_{11}, y_{21}, y_{31}은 처리집단1의 평균 $\bar{y}_1$과 (약간의) 오차항 e_{i1}의 합으로 각각 설명된다. 다시 말해, $y_{i1} = \bar{y}_1 + e_{i1}$이다.
- 처리집단1의 평균 $\bar{y}_1$이 곧바로 처리효과가 되는 것은 아니다. 처리효과는 처리집단1의 평균 $\bar{y}_1$과 총평균 $\bar{y}$의 차이로 나타난다(Why?). 즉, $\alpha_1 = \bar{y}_1 - \bar{y}$이다.
- 그러므로 $y_{i1} = \bar{y}_1 + e_{i1} = \bar{y} + \alpha_1 + e_{i1}$이다.

처리집단1	처리집단2	처리집단3	
$y_{11} = 2$	$y_{12} = 3$	$y_{13} = 15$	
$y_{21} = 2$	$y_{22} = 2$	$y_{23} = 13$	
$y_{31} = 5$	$y_{32} = 1$	$y_{33} = 20$	
$\bar{y}_1 = 3$	$\bar{y}_2 = 2$	$\bar{y}_3 = 16$	$\bar{y} = 7$

일원분산분석

- 우리가 보고 있는 결과인 y_{ij}은 사실 (처리와는 별개로 우리 모두가 나름대로 가진) 일종의 출발선(baseline)이라고 할 수 있는 총평균 μ 뿐만 아니라, 그룹 간(between) 처리효과 α_j와 그룹 내(within) 개별적인 오차 e_{ij}로부터도 영향을 받아 나타난 것이다!

$$y_{ij} = \mu + \alpha_j + e_{ij}$$

- 이제 위 식을 살짝 변경하면 일원분산분석의 핵심을 모두 파악할 수 있다.

$$\begin{aligned} Y_{ij} - \mu &= \alpha_j + e_{ij} \\ &= (\mu_j - \mu) + (Y_{ij} - \mu_j) \end{aligned}$$

- 위 식에 제곱합(sum of squares; SS)을 해도 아래와 같이 성립한다는 것은 (우리는 하지 않지만!) 수학적으로 증명할 수 있다.

$$\sum (Y_{ij} - \mu)^2 = \sum (\mu_j - \mu)^2 + \sum (Y_{ij} - \mu_j)^2$$

$$SS_{total} = SS_{between} + SS_{within}$$

일원분산분석

- 제곱합 SS는 적절한 자유도인 $n-1$와 $k-1$ 그리고 $n-k$로 나누어 평균제곱(mean squares; MS)이라고 부른다.

$$\frac{\sum(Y_{ij}-\mu)^2}{(n-1)} = \frac{\sum(\mu_j-\mu)^2}{(k-1)} + \frac{\sum(Y_{ij}-\mu_j)^2}{(n-k)}$$

$$MS_{total} = MS_{between} + MS_{within}$$

- 각각의 자유도 계산원리도 무척 단순하다(Why?). 여기서 k은 그룹의 숫자이다.

$$(n-1) = (k-1) + (n-k)$$

- 그리고 이것은 각각 "분산의 불편추정량"이 된다(Why?).

$$\sigma^2_{total} = \sigma^2_{between} + \sigma^2_{within}$$

일원분산분석

일원분산분석의 검정통계량은 F 값이다.

- 일원분산분석은 F 분포를 사용하며, 그 근본 원리는 (앞서 설명한) 두 모집단 분산 비율에 대한 가설검정과 완전히 같다.

$$
\begin{aligned}
F_{(k-1,n-k)} &= \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)} \\
&= \frac{MS_{between}}{MS_{within}} \\
&= \frac{\sigma^2_{between}}{\sigma^2_{within}}
\end{aligned}
$$

- 분자는 실험처리에 의한 효과(treatment effect)로 볼 수 있고, 분모는 무작위오차(random error)에 지나지 않는다(Why?).

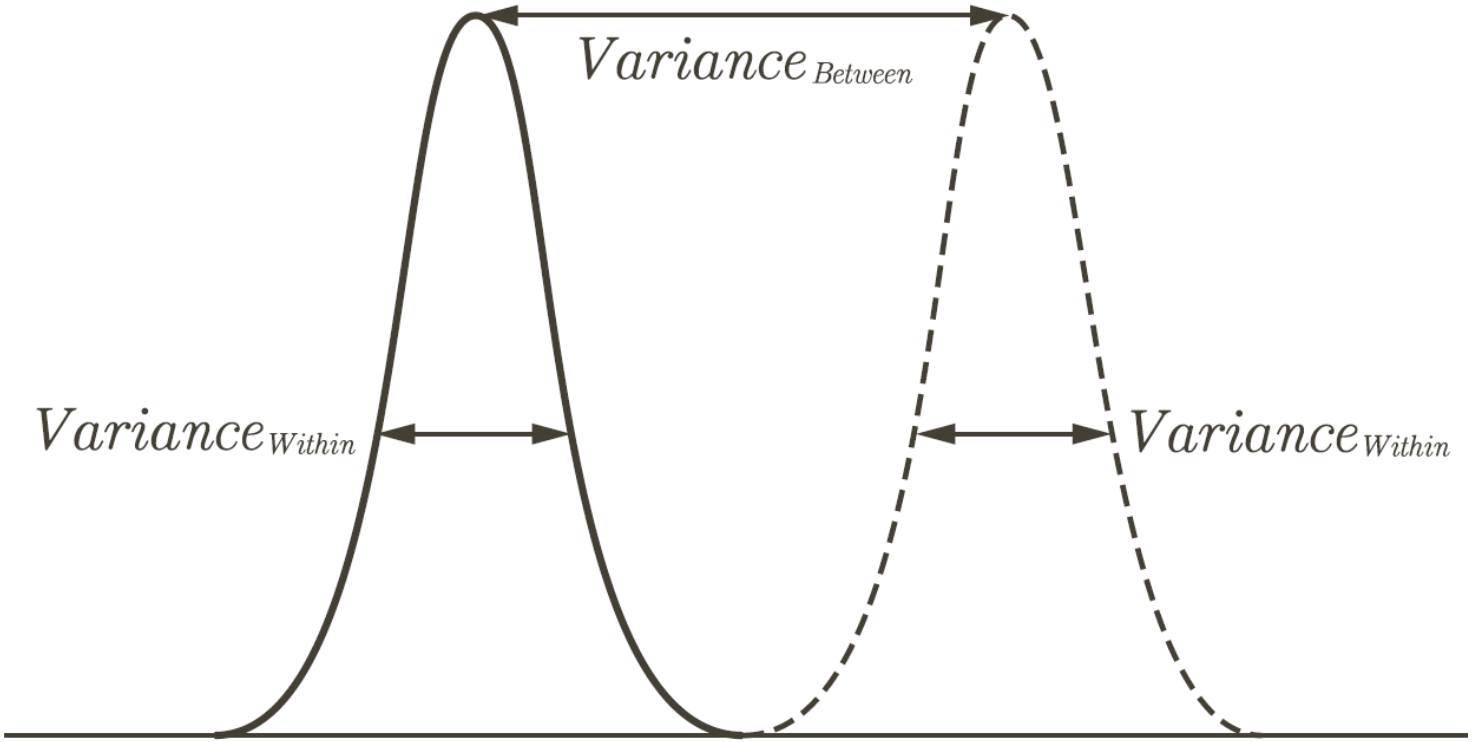
일원분산분석

집단 간 분산이 클수록 그리고 집단 내 분산이 작을수록 F 값은 커진다.

- 수리적으로 따지고 들어갈 필요조차 없다. 사실 F 값이 갖는 의미는 상식적으로 너무나 명확하다!
- 두 표본이 주어졌을 때, 표본 간에는 차이가 크고 표본 내에는 차이가 작다면, 두 모집단은 서로 다른 것이다.
- 두 표본이 주어졌을 때, 표본 간에는 차이가 작고 표본 내에는 차이가 크다면, 두 모집단은 서로 다르다고 할 수 없다.
- 따라서 F 값이 크다는 것은 두 모집단은 서로 다름을 시사한다.

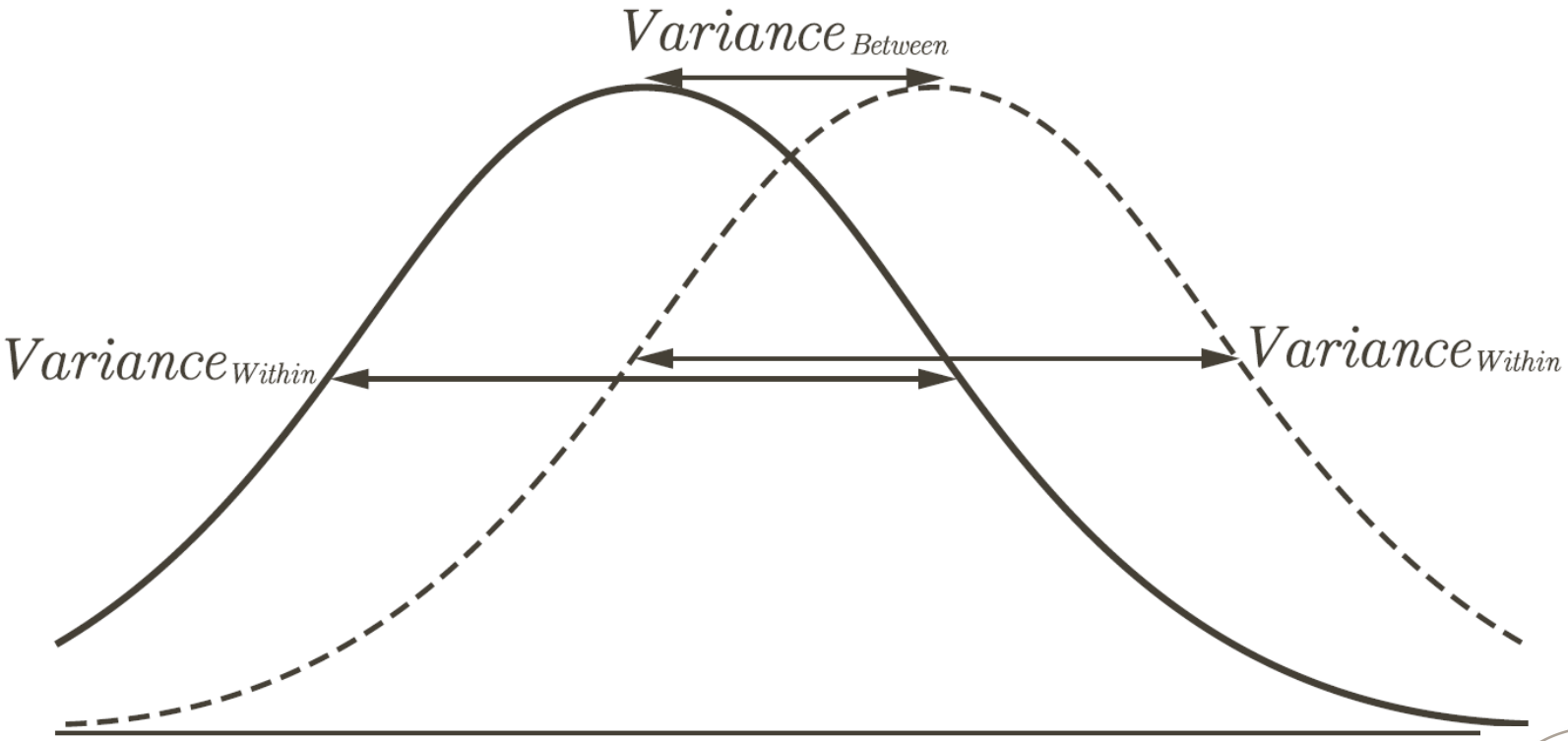
일원분산분석

- 집단 간(between) 차이가 크고 집단 내(within) 차이가 작은 경우



일원분산분석

- 집단 간(between) 차이가 작고 집단 내(within) 차이가 큰 경우



일원분산분석

- 계산을 진행하면서 분산분석표(ANOVA table) 안에 각각의 숫자를 채워넣으면 된다. 아! 물론 계산은 컴퓨터가 한다.

Variance	SS	df	MS	F
Between groups	$SS_{between}$	$(k-1)$	$MS_{between}$	F
Within groups	SS_{within}	$(n-k)$	MS_{within}	
Total	SS_{total}	$(n-1)$	MS_{total}	

- 실제 표에서는 F 옆에 유의확률(p-value)이나 임계값(critical value) 등도 추가해서 써넣는 경우도 있다.

일원분산분석

예제 1. 다음 분산분석표의 빈 칸을 채우시오.

Variance	SS	df	MS	F	p
Between groups	9.214	3	3.071	(나)	0.010
Within groups	(가)	135	0.775		
Total	113.827	138			

일원분산분석

- 이 문제는 사회조사분석사 2급 2017년 2회에 기출문제로 출제된 바 있다.

Variance	SS	df	MS	F	p
Between groups	9.214	3	3.071	(나)	0.010
Within groups	(가)	135	0.775		
Total	113.827	138			

- (가)는 SS_{within} 이므로 $9.214 + (가) = 113.827$로 알 수 있다.
- 또는 $(가)/135 = 0.775$로도 알 수 있다.
- (나)는 F 이므로 $3.071/0.775$로 알 수 있다.

일원분산분석

일원분산분석의 가설 설정은 다소 주의를 요한다.

- 귀무가설은 "모든 그룹에 걸쳐 평균이 같다"이다. 예를 들면 다음과 같다.

  (1) 인지된 기후변동의 심각성 점수의 평균에 최종학력별 차이는 없다.
  (2) 출신지역별로 정치적 보수주의 점수 차이가 없다.
  (2) 기업내 직급 수준에 따른 인적자원개발 프로그램 만족도 차이는 없다.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

- 그 대립가설은 "적어도 하나의 그룹에서 평균값이 다르다"이다.

$$H_a : (\mu_1 \neq \mu_2) \text{ or } (\mu_1 \neq \mu_3) \text{ or } ... \text{ or } (\mu_{j-1} \neq \mu_j)$$

일원분산분석

- "모든 그룹에 있어 평균값이 다르다($H_a : \mu_1 \neq \mu_2 \neq \mu_3 \neq ... \neq \mu_j$)"가 아님에 주의할 것(Why?)!
- 당연히 "최종학력이 높아질수록 인지된 기후변동의 심각성 점수도 함께 높아진다"와 같은 해석은 본래 일원분산분석으로 할 수 없는 과잉해석이 된다.
- 일원분산분석은 실제 분석상 분산의 비율을 비교하고 있음에도 불구하고, 가설 설정은 평균에 대해 이루어진다는 점도 주의해야 한다.

일원분산분석

두 개의 자유도와 계산된 $F_{(k-1,n-k)}$ 값을 가지고 유의성 검정을 수행한다.

- 귀무가설이 옳다는 전제 아래 두 모집단에서 무한히 계속 표본을 뽑아 그 분산 비율 s_1^2/s_2^2의 표집분포를 그린다.
- 이 표집분포는 (두 개의 자유도에 근거하여 모양이 정해지는) F 분포를 따른다.
- 여기서 검정통계량 F 값의 위치와 색칠공부 영역을 가늠한다.
- 만일 검정통계량 F 값이 충분히 커서 오른쪽 끄트머리의 유의확률(p-value)이 α보다 작으면 $1-\alpha$ 신뢰수준에서 귀무가설을 기각한다.
- (앞서 배웠듯이) F 분포의 "곡선 밑 면적"을 계산하는 엑셀 함수는 F.DIST(x, deg_freedom1, deg_freedom2, TRUE)이다.

일원분산분석의 활용

일원분산분석의 활용

예제 2. 새우는 노바 제천에서 수많은 커피체인점을 열어 떼돈을 벌어들이고 있다. 이제부터 과학적으로 더 많은 돈을 벌어야겠다고 결심한 새우는 커피 원두를 각각 브라질(1), 콜롬비아(2), 과테말라(3), 에티오피아(4) 네 곳에서 수입하여 자신이 소유한 커피체인점 여섯 곳에 분배하였다. 점원들에게는 고객이 "아메리카노 주세요" 했을 때, 네 가지 원두 중 하나를 랜덤하게 사용하도록 교육하였다. 한 달 후, 이제 새우는 원두 수입처가 커피의 판매량과 관계가 있는가를 알아보기 위해 coffee.csv 자료를 만들었다. 커피 판매량이 커피 원두에 따라 다르다고 할 수 있는지 유의수준 5%에서 검정하시오.

일원분산분석의 활용

- 자료를 잘 살펴보아 구조를 파악해보자. 확실히 아까 설명한 일원분산분석의 자료구조를 닮아있다!
- 아까 설명한 자료구조 꼴처럼 자료가 정리되어 있어야만 엑셀에서 일원분산분석을 수행할 수 있다(단 행과 열이 뒤바뀌는 것까지는 가능하다).
- 좀 더 전문적인 통계분석 패키지라면 자료구조와 상관없이 일원분산분석을 자유롭게 수행할 수 있다.

일원분산분석의 활용

- 엑셀에서 [데이터]-[데이터 분석]을 따라 들어가 “분산 분석: 일원 배치법”을 선택한다.
- 자료를 적절하게 하이라이트한 다음, “첫째 행 이름표 사용” 및 “유의 수준”을 적절히 선택 또는 입력한다.
- 분산분석표를 보고 해석한다.
- 제곱합, 자유도, 제곱 평균은 각각 어떻게 계산되었나?
- F 값과 유의 확률(p-value)은 어떻게 계산되었나?
- 분석의 결론은 무엇인가? 어떤 결론을 내려선 안되는가?

일원분산분석의 활용

양적 연구에서 일원분산분석은 주로 두 가지 용도를 갖는다.

- 첫번째는 표본에 관한 기술통계(descriptive statistics)를 제시하는 부분이고, 두번째는 회귀분석(regression analysis)에서 모형 적합도(goodness-of-fit) 평가 부분이다(이것은 나중에 배운다).
- 기술통계 분석의 맥락에서 보면, 연구자는 자신의 표본 안의 핵심이 되는 "범주형" 변수를 요인(factor)으로 삼고 다른 여러 "연속형" 변수를 결과(outcome)로 삼아 일원분산분석을 수행할 수 있다.
- (앞서 설명하였듯) 범주형 변수는 명목 (내지 서열) 척도로 측정된 것이다. 가령 최종학력, 출신지역, 지지하는 정당, 고용상태(이희정 2018) 등을 생각해 볼 수 있다.

이희정. 2018. "청년층 계층인식 변화가 공정성 인식에 미치는 영향 분석." 한국사회학 52(3): 119–164.

일원분산분석의 활용

		Average of effort reward fairness	F	Prob.>F
All		2.981		
Sex	Male	2.975	0.10	0.7469
	Female	2.987		
Co-residency with parents	Not living together	3.008	1.02	0.3136
	Living together	2.969		
Economic Independence	Independent	3.031	7.47	0.0063
	Dependent	2.935		
Marital status	Married	3.120	24.65	0.0000
	Single	2.927		
Employment	Regular worker	3.067	9.36	0.0001
	Self-employed	2.866		
	Unemployed	2.938		
Residence	Seoul metropolitan region	2.851	56.29	0.0000
	Others	3.110		
Home ownership	Owner	3.033	11.09	0.0009
	Others	2.916		

일원분산분석의 활용

일원분산분석을 목적에 맞게 사용해야 한다.

- 평균을 비교하려고 할 때, 그룹이 2개라면 t 검정을 사용하고, 2개 이상이라면 일원분산분석을 사용한다.
- 그럼 만일 그룹이 2개만 주어졌을 때 (t 검정 대신) 일원분산분석을 수행하면 어떤 결과를 가져올까?
- 본래 일원분산분석의 귀무가설은 "모든 집단에 걸쳐 평균값이 동일하다 ($H_0 : \mu_1 = \cdots = \mu_j$)"였다.
- 그룹이 두 개만 주어진 귀무가설의 경우 "두 집단에 걸쳐 평균값이 동일하다 ($H_0 : \mu_1 = \mu_2$)"로 축소된다.
- 이것은 t 검정의 귀무가설($H_0 : \mu_1 - \mu_2 = 0$)과 본질적으로 같다.
- 실제로 F 값과 t 값에는 밀접한 관계가 있다(증명 생략).

$$\sqrt{F} = |t| \quad (\text{또는 } F = t^2)$$

일원분산분석의 활용

- 반대로 2개 이상의 그룹에 대해 일원분산분석 대신 t 검정을 여러번 하면 안될까?
- 결론만 말하자면 (1) 굉장히 불편하고 혼란스러울 뿐 아니라, (2) 추정상 오류를 저지르게 될 위험이 극단적으로 커지므로 안된다.
- 첫째, t 검정을 아주 여러 번 수행하고 비교해야 하는 부담이 있다. 예컨대 겨우 5개의 모집단을 비교하기 위해서 t 검정을 10번이나 수행해야 한다(Why?). 게다가 이 횟수는 기하급수적으로 증가한다.
- 둘째, 이 10번의 t 검정을 수행하는 과정에서 최소 1번 이상 오류가 나타날 가능성은 급격히 증가한다. (이항분포로 계산하면 알 수 있듯) 예컨대 5% 유의확률이라면 1회 이상의 오류 확률은 40%나 된다(Why?).

일원분산분석의 활용

일원분산분석은 사실 몇 가지 가정에 입각해 있다.

- 다음과 같이 일원분산분석의 수리모형이 주어져 있을 때, 오차항 e_{ij}에 관해 몇 가지 가정이 필요하다.

$$Y_{ij} = \mu + \alpha_j + e_{ij}$$

(1) 선형성 가정: $E(e_{ij}|\alpha_j) = 0$
(2) 독립성 가정: $e_{ij} \perp\!\!\!\perp e_{i'j'}|\alpha_j \quad \forall_{i,j}\, i \neq i',\, j \neq j'$
(3) 등분산성 가정: $Var(e_{11}|\alpha_j) = Var(e_{21}|\alpha_j) = ... = Var(e_{ij}|\alpha_j)$
(4) 정규성 가정: $e_{ij} \sim N(\mu, \sigma^2)$

- 사조사 기출 문제에 대비하는 심정으로 외워두자.