사회통계연습

χ^2 분포, F 분포 그리고 ANOVA

김현우, PhD[1]

[1] 충북대학교 사회학과 부교수

진행 순서

1. 일원분산분석의 논리
2. 일원분산분석의 절차

일원분산분석의 논리

일원분산분석의 논리

분산분석으로 표본(=변수)의 분산에 관한 가설 검정을 수행한다.

- Ronald Fisher는 F 분포를 만들고 그것을 절묘하게 활용하는 분산분석(Analysis of Variance; ANOVA)도 개발했다.
- 일원분산분석(one-way ANOVA)은 분산분석에 속한 여러 기법들의 가장 기초가 된다. 분산분석의 기본적인 개념을 학습하는데 유용할 뿐 아니라, 실제로도 여러가지 맥락에서 종종 사용된다.
- 다만 (심리학을 제외한) 다른 사회과학 분야에서는 실험자료(experimental data) 대신 관찰자료(observational data)가 더 폭넓게 사용된다(다시 말해, 분산분석이 폭넓게 사용되기 어려운 편이다).

일원분산분석의 논리

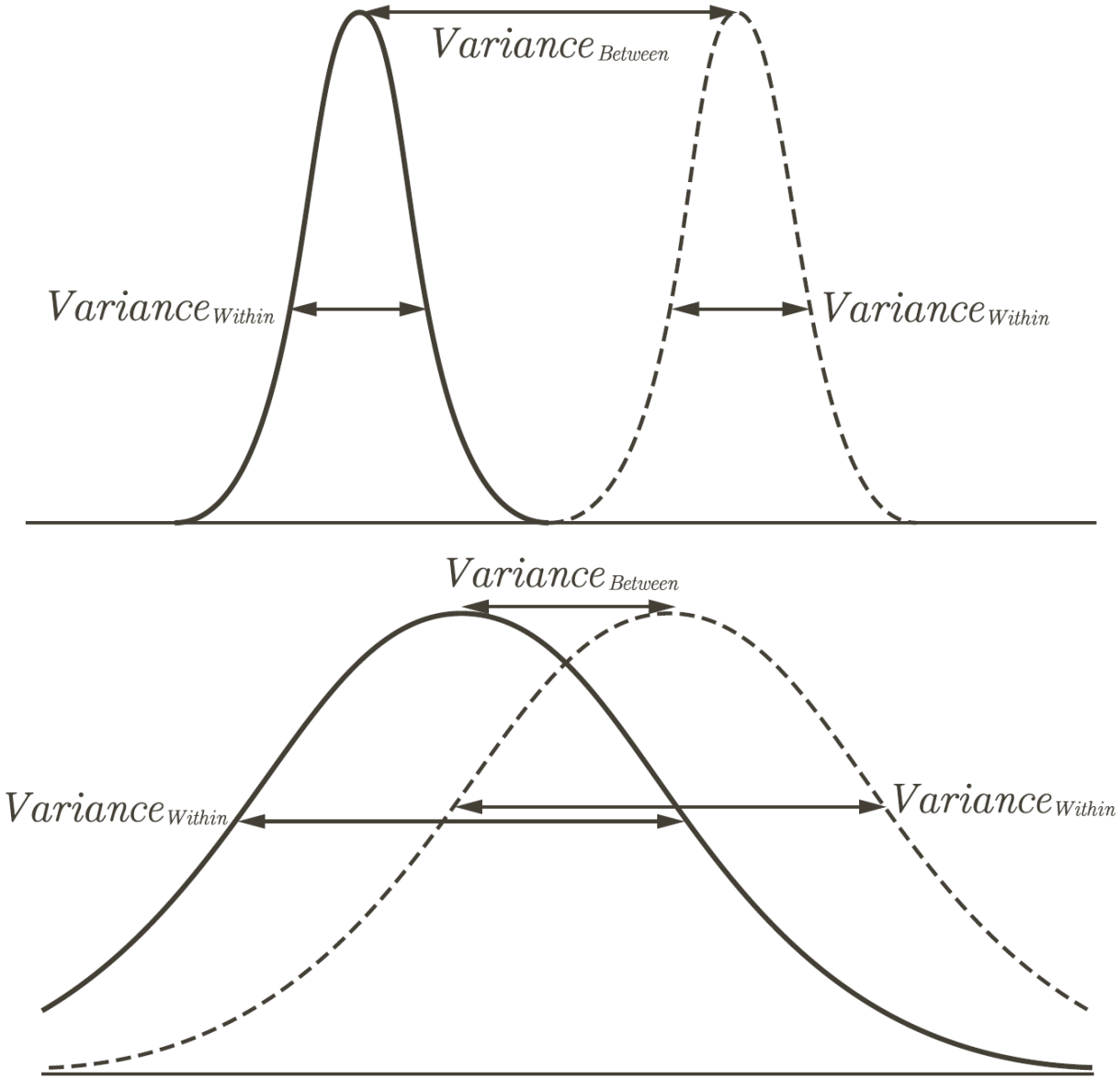
일원분산분석은 F 검정의 논리를 사용한다.

- 집단 간 분산(between-group)이 클수록 그리고 집단 내 분산(within-group)이 작을수록 F 값은 커진다.

$$F_{(n-1,n-k)} = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/(k-1)}{SS_{within}/(n-k)}$$

- 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.

일원분산분석의 논리


Variance Between
Variance Within
Variance Within
Variance Between
Variance Within
Variance Within

일원분산분석의 논리

그런데 왜 F 값은 분산의 비율일까?

- 우리는 모집단에서 표본을 무한히 계속 뽑고 그 표본분산 s^2를 계속 구해 분산의 표집분포를 구축한다고 상상해보자.
- 하지만 아쉽게도 이러한 표본분산의 표집분포 s^2에 관한 직접적인 이론 분포는 없다!
- 대신 "표본분산 s^2와 모분산 σ^2 간의 비율"에 관한 이론적 확률분포인 χ^2 분포 (chi-square distribution)가 있다.

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

일원분산분석의 논리

- 단일모집단의 분산에 관한 가설 검정은 χ^2 분포를 이용할 수 있다. 하지만 (우리가 하려는 것처럼) 두 모집단의 분산의 비율에 관한 가설 검정이라면 F 분포를 이용해야 한다(교재 190).
- F 분포의 확률밀도함수는 다음과 같이 정의된다.

$$F_{(n_1-1,n_2-1)} = \frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)}$$

- 조금만 생각해보면 F 분포는 각각의 단일모집단의 분산을 설명하는 χ^2 의 비율임을 쉽게 파악할 수 있다(Why?).
- 두 표본의 분산의 차(difference)를 구하려는 것이 아니라 비(ratio)를 구하려는 것이므로, 두 χ^2 값이 서로 비슷할수록 F 값은 (0이 아니라) 1에 접근하리라 예상할 수 있다.
- 그러므로 일원분산분석은 χ^2 검정에서 F 검정을 거치고나서 겨우 이해할 수 있는 기법이다.

일원분산분석의 논리

- 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.

일원분산분석의 논리

일원분산분석의 가설 설정은 다소 주의를 요한다.

- 귀무가설은 모든 그룹에 걸쳐 평균이 같다는 내용이 된다. 예를 들면 다음과 같다: "출신지역 별로 연평균 소득 평균에는 차이가 없다."

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

- 그 대립가설은 "적어도 하나의 출신지역에서 연평균 소득이 다른 지역과는 다르다"가 된다.

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

- "모든 그룹에 있어 평균값이 다르다($H_a : \mu_1 \neq \mu_2 \neq \mu_3 = ... \neq \mu_j$)"가 아님에 주의할 것(Why?)!
- 일원분산분석은 실제 분석상 분산의 비율을 비교하고 있음에도 불구하고, 가설 설정은 평균에 대해 이루어진다는 점도 주의해야 한다.

일원분산분석의 절차

일원분산분석의 절차

일원분산분석 절차는 굉장히 정형화되어 있다.

- 분석을 시작하기 전에 사용할 변수들이 무엇이고, 그들의 측도는 무엇인지 파악한다.
- 결측치와 극단치(outliers)의 존재를 파악하기 위해 먼저 기술통계 및 시각화를 수행하는 것이 좋다.
- SPSS에서는 [분석]-[평균 비교]-[일원배치 분산분석]을 선택하여 일원분산분석을 수행한다. 이때 "기술통계"를 옵션에서 선택하는 쪽이 좋다.
- Jamovi에서는 [분석]-[분산 분석]-[일원 분산 분석]을 선택한다.
- 분산분석표를 보고 해석할 수 있어야 한다. 제곱합, 자유도, 제곱 평균은 각각 어떻게 계산되었는지 이해한다.
- F 값과 유의 확률(p-value)은 어떻게 계산되었는지 이해한다.
- 분석의 결론은 무엇인지, 어떤 결론을 내려선 안되는지 판단한다.

일원분산분석의 절차

일원분산분석의 전형적인 표와 시각화 기법이 있다.

- 이희정(2018)과 이미숙(2012)의 논문에 실린 표를 살펴보면서 의미를 해석해보자.
- 좋은 표를 최대한 흉내내는 연습이 필요하다!
- 시각화 기법으로 무엇이 적절할까? 반드시 시각화해야 하는 것은 아니지만 필요하다고 생각하면 넣을 수 있다.

일원분산분석의 절차

〈표 3〉 노인인구의 결혼지위 및 결혼만족도 유형에 따른 우울증세 차이

유형 구분			빈도(%)	CES-D 평균	분산분석
결혼지위 (N=4,040)	전체 (N=4,012)	혼인	2,588(64.5)	17.86	F=33.3 (p<.001)
		별거	22(0.5)	18.86	
		사별/실종/이산가족	1,362(33.9)	19.93	
		이혼	31(0.8)	21.71	
		미혼	9(0.2)	21.11	
	여성 (N=2,333)	혼인	1,084(46.5)	18.60	F=9.29 (p<.001)
		별거	16(0.7)	19.97	
		사별/실종/이산가족	1,211(51.9)	19.97	
		이혼	16(0.7)	22.25	
		미혼	6(0.3)	22.33	
	남성 (N=1,679)	혼인	1,504(89.6)	17.33	F=8.37 (p<.001)
		별거	6(0.4)	16.67	
		사별/실종/이산가족	151(9.0)	19.59	
		이혼	15(0.9)	21.13	
		미혼	3(0.1)	18.67	

이미숙. 2012. "노인인구의 결혼관계와 우울증세: 결혼지위와 결혼만족도를 중심으로." 『한국사회학』 46(4): 176-204.

일원분산분석의 절차

〈Table 7〉 Analysis of variance on perception toward justice

		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		

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

일원분산분석의 절차

연습 1. citytemp.sav에서 난방도일(heatdd)과 냉방도일(cooldd)이 센서스 구분(region)에 따라 다른지 여부를 검정하시오. 표와 시각화를 통해 그 결과를 요약하시오.

일원분산분석의 절차

	Heating degree days		Cooling degree days	
	평균	표준편차	평균	표준편차
NE	5803	743	722	238
N Cntrl	6446	1006	822	311
South	2518	1390	2327	888
West	3169	1948	973	880
전체	4426	2200	1240	938
F (p>F)	482.482 (<.001)		299.018 (<.001)