계량분석
One-Way ANOVA in Practice

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 17, 2024

진행 순서

1. 일원분산분석의 실제 활용
2. 연습문제

일원분산분석의 실제 활용

일원분산분석의 실제 활용

경험적 연구에서 일원분산분석은 크게 두 부분에서 주로 활용된다.

- 첫째로 표본에 관한 기술통계(descriptive statistics)를 제시할 때 사용될 수 있고, 두번째로 회귀분석(regression analysis)에서 모형 적합도(goodness-of-fit) 지표 중 하나로 사용될 수 있다.
- 기술통계로 제시한다면, 표본 안의 핵심이 되는 범주형 관심변수 내지 종속변수에 따라 다른 여러 변수들이 어떻게 다른지 일원분산분석을 통해 보여줄 수 있다.
- 이때 범주형 변수는 명목형 내지 순서형 척도로 측정된 것이며 예컨대 최종학력, 출신지역, 지지하는 정당 등을 생각해 볼 수 있다.

일원분산분석의 실제 활용

〈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 실제 활용

〈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.

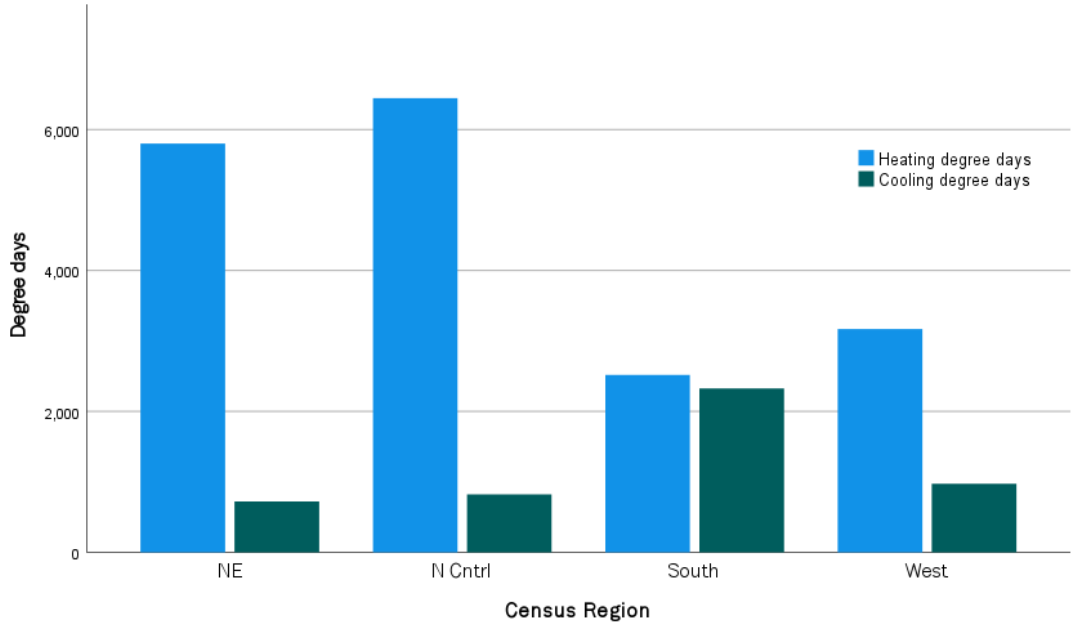
일원분산분석의 실제 활용

일원분산분석의 전형적인 표와 시각화 기법이 있다.

- citytemp 데이터에서 난방도일(heatdd)과 냉방도일(cooldd)이 센서스 지역구획(region)에 따라 다른지 여부를 검정해보자.
- 표와 시각화를 통해 그 결과를 적절히 요약해보자.
- 좋은 표를 최대한 흉내내는 연습이 필요하다(당연히 숙제로 나간다)!
- 시각화 기법으로 무엇이 적절할까? 반드시 시각화해야 하는 것은 아니지만 필요하다고 생각하면 넣을 수 있다.

일원분산분석의 실제 활용

	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)	



일원분산분석의 실제 활용

일원분산분석은 회귀분석 전체 계수의 유의성 검정에서도 사용된다.

- 이 경우 귀무가설은 "모든 회귀계수들이 0이다"로 만일 이 귀무가설을 기각하지 못한다면 모델에 포함된 어떠한 독립변수 X로도 종속변수 Y를 의미있게 설명하지 못함을 의미한다(Why?).
- 당연히 이 경우에는 회귀모형을 처음부터 다시 만들어야 한다.
- 회귀분석의 맥락에서 대립가설은 "적어도 하나 이상의 회귀계수는 0이 아니다" 임에 주의할 것.
- 우리는 나중에 회귀분석을 배우면서 일원분산분석이 회귀분석의 맥락에서도 다시 한 번 쓰이게 됨을 확인하게 된다.

일원분산분석의 실제 활용

t 검정과 일원분산분석에는 잘 알려진 연관성이 있다.

- t 검정은 두 모집단의 평균을 비교하고, 일원분산분석은 여러 모집단에 걸친 분산의 비율을 비교한다.
- 만일 집단이 두 개만 주어졌을 때 일원분산분석을 수행하면 어떤 결과를 가져올까?
- 이 경우 일원분산분석의 귀무가설은 "모든 집단에 걸쳐 평균값이 동일하다"였으므로 이는 다시 "두 집단에 걸쳐 평균값이 동일하다"로 축소된다.
- 즉 t 검정과 같은 것이 된다. 실제로 F 값과 t 값에는 다음과 같은 관계가 있다.

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

- auto.dta에서 수입품 여부(foreign)에 따라 가격(price) 차이가 있는지 여부를 t 검정과 일원분산분석으로 함께 검정해보고, 그 차이를 살펴보자.

일원분산분석의 실제 활용

- 그렇다면 반대로 생각해서, 집단이 여러 개 있을 때 구태여 일원분산분석 대신 t 검정을 여러 번 하면 안될까?
- 결론만 말하자면 (1) 굉장히 불편하고 혼란스러울 뿐 아니라, (2) 추정상의 오류를 저지르게 될 위험이 극단적으로 커지므로 권할 수 없다.
- 먼저 t 검정을 아주 여러 번 수행하고 비교해야 하는 부담이 있다.
- 예컨대 겨우 5개의 모집단을 비교하기 위해서 t 검정을 10번이나 수행해야 한다 (Why?).
- 이것은 기하급수적으로 증가하여 6개의 모집단을 비교하기 위해서는 t 검정을 15번이나 수행해야 한다(Why?).

일원분산분석의 실제 활용

- 게다가 이 10번의 t 검정을 수행하는 과정에서 최소 1번 이상 오류가 나타날 가능성 또한 급격히 증가한다.
- 예컨대 5% 유의확률이라면 1회 이상의 오류 확률은 약 40%나 된다(Why?).
- 이항분포(binomial distribution)를 통해 이 확률분포를 계산할 수 있다. 즉 발생확률 p가 0.05인데 10번의 시행 중 사건이 전혀 발생하지 않을 확률분포는 다음과 같다.

$$1 - \binom{n}{k} p^k (1-p)^{n-k} = 1 - \binom{10}{0} \cdot 0.05^0 \cdot (1-0.05)^{10-0}$$

- 이 사실을 간단히 Stata에서 계산해보자.

연습문제

연습문제

연습 1. 서울시민 여성혐오 및 여성정책에 대한 인식조사 자료에서 성별(SQ2) 및 연령대별(SQ3_1)로 총 10개의 범주를 생성하시오(e.g., “여성(30-39)”, “남성(15-19)” 등). 또한 다양한 여성혐오 표현에 대한 관대함(A3)을 모두 합산하여 그 관대함의 정도를 나타내는 합성지수를 계산하시오. 여성혐오 표현에 대한 관대함이 성별-연령대별 범주에 따라 상이한지 살펴보기 위한 유의성 검정을 수행하고, 그 차이를 시각화하고 차이를 간단히 해설하시오.