계량분석
Comparing Variances

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 18, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 단일모집단 분산에 대한 가설검정
2. 두 모집단 분산에 대한 가설검정
3. 일원분산분석
4. 일원분산분석의 실제 활용

단일모집단 분산에 대한 가설검정

단일모집단 분산에 대한 가설검정

어떤 모집단의 분산(σ^2)을 추리하고 싶은 여러 가지 동기가 있다.

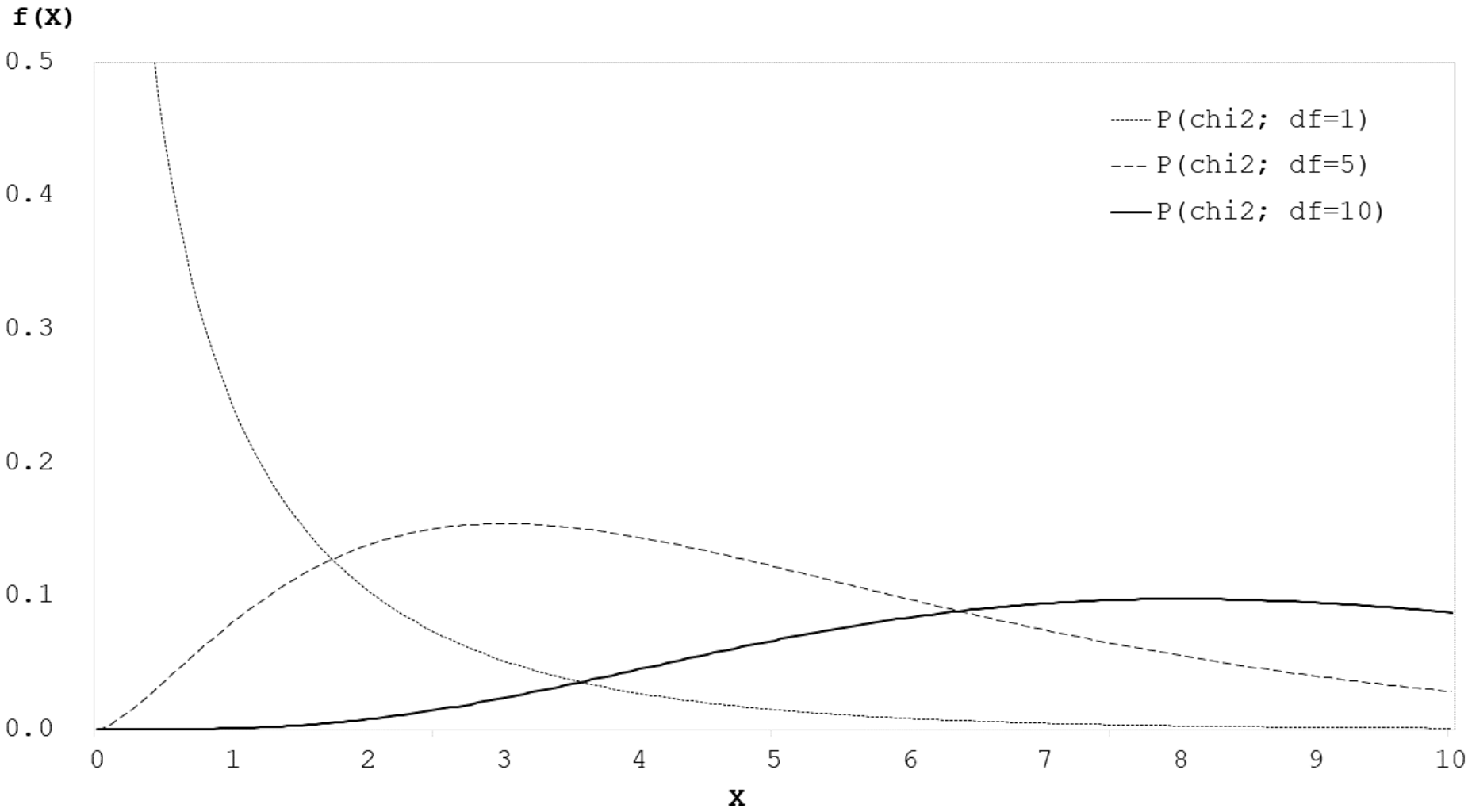
- 예컨대 표본 안에서 수집된 소득의 분산(variance)은 소득 불균등을 의미한다(사실 사회계층 및 불평등 연구에서는 더 정교한 소득 불균등 지표가 선호된다). 금융상품의 수익금의 분산은 해당 상품의 리스크(risk)를 의미한다. 제조된 음료수 용량의 분산은 들쭉날쭉한 품질 상태를 의미한다.
- 주어진 모집단에서 일정한 크기(n)의 표본을 무한히 많이 뽑아 그 표본분산(sample variances)들로 (가상적인) 표집분포(sampling distribution)를 그린다면, 이는 자유도(degree of freedom)가 n-1인 χ^2 분포(chi-square distribution)를 따른다.

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

- χ^2 분포는 비대칭적이고, 오른쪽으로 꼬리가 길며, 항상 양수값을 갖는다.
- χ^2 분포의 모양은 오로지 자유도에 의해서만 결정된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

단일모집단 분산에 대한 가설검정



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

단일모집단 분산에 대한 가설검정

앞서 배운 바와 마찬가지로 유의성 수준(significance level)에 따라 구체적인 영가설 기각의 기준을 세울 수 있다.

- 1,000명의 학생들에 대해 일괄적으로 시험을 실시하였다. 그리고 50명의 학생을 임의 표본(random sample)으로 추출하였다.
- "모집단의 표준편차는 15점이다($\sigma = 15$)"라는 영가설을 세우고, 이에 따라 50개의 표본으로 이루어진 (가상적인) 표집분포로 χ^2_{49} 분포를 그릴 수 있다.
- 표본에서 계산된 표준편차가 20점이라고 하자. 이 경우 χ^2 값은 87.11이다 ($= 49 \cdot 20^2/15^2$)이다.
- 5% 유의수준(=95% 신뢰수준)을 기준일 때, Stata에서는 **(1-chi2(49, 87.11))*2**를 계산하여 유의확률(p-value)을 구할 수 있다(Why?). 답은 약 .0013이다.
- p-value가 0.01보다 작으므로 영가설을 1% 유의수준에서 기각한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

단일모집단 분산에 대한 가설검정

Stata에서는 **sdtest** 명령어로 단일모집단의 분산에 관한 가설 검정을 할 수 있다(**[Stata 코드]** 참고).

- 다만 이것은 분산을 비교하는게 아니라 표준편차를 비교한다는 점에 주의하자.
- 먼저 social.dta 파일을 열고 socialself 변수의 요약통계량을 살펴보자.
- "socialself의 표준편차가 5.1이다."라는 영가설을 테스트해보자.
- "socialself의 표준편차가 5.1보다 작거나 같다" 라는 영가설을 세웠다. 이를 테스트해보자.
- "socialself의 표준편차가 5.1이다 크거나 같다" 라는 영가설을 세웠다. 이를 테스트해보자.

결과표를 꼼꼼히 들여다보자.

- χ^2 값과 자유도(degree of freedom)은 어떻게 계산되었는가?
- 대립가설 별로 유의확률(p-value)은 어떻게 계산되었는가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 모집단 분산에 대한 가설검정

두 모집단 분산에 대한 가설검정

분산비교(variance comparison)는 보통 F 분포(F distribution)에 기반하여 수행된다.

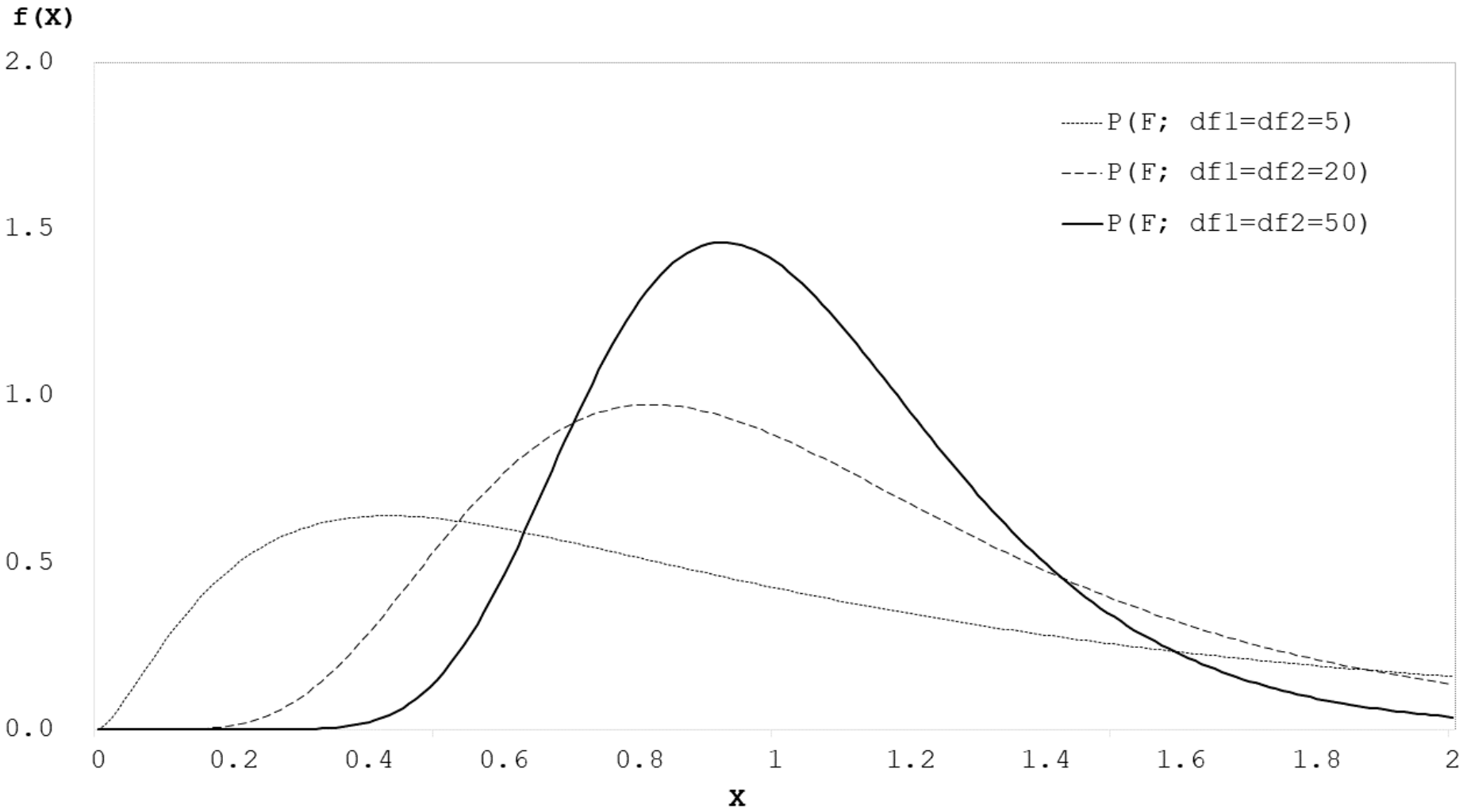
- "단일모집단의 분산"에 관한 가설 검정이 χ^2 분포를 이용하는 반면, "두 모집단의 분산의 비율"에 관한 가설 검정은 F 분포를 이용한다.
- F 분포의 확률밀도함수(PDF)는 다음과 같이 정의된다:

$$\mathrm{F}_{(\mathrm{n}_1-1,\mathrm{n}_2-1)} = \frac{\chi_1^2/(\mathrm{n}_1-1)}{\chi_2^2/(\mathrm{n}_2-1)}$$

- 상식적으로 조금만 생각해보면 F 분포는 각각의 "단일모집단의 분산"을 설명하는 χ^2의 비율임을 쉽게 파악할 수 있다. 다만 자유도를 사용하여 약간의 보정이 이루어졌을 뿐이다. 심지어 $\mathrm{n}_1 = \mathrm{n}_2$라면 그마저도 필요없다.
- F 분포는 두 개의 자유도 $(\mathrm{n}_1 - 1)$와 $(\mathrm{n}_2 - 1)$를 파라미터(parameter)로 받아 그 형태가 결정된다. 이것도 χ^2가 하나의 자유도 $(\mathrm{n} - 1)$를 파라미터로 받아 그 형태가 결정되는 것과 매우 유사하다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

단일모집단 분산에 대한 가설검정



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 모집단 분산에 대한 가설검정

일단 χ^2 분포 자체에 익숙해져야 F 분포에 대해서도 쉽게 이해할 수 있다.

- 우리는 두 표본의 분산의 차(difference)를 구하려는 것이 아니라 비(ratio)을 구하려는 것이므로 두 χ^2 값이 서로 비슷할수록 F 값은 0이 아니라 1에 접근하리라 예상할 수 있다.
- F 분포는 두 개의 χ^2 분포의 비(ratio)로 구성되므로 0보다 작을 수 없다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

두 모집단 분산에 대한 가설검정

이제 Stata에서 두 개의 독립표본(independent samples) 데이터를 가지고 분산비교를 연습해보자(**[Stata 코드]** 참고).

- 다시 social_independent.dta 파일을 열고 wave 별로 socialself 변수의 요약통계량을 살펴보자.
- wave 전후로 socialself의 분산(=표준편차^2)이 달라졌는지 **sdtest** 명령어로 살펴보고, "두 변수의 표준편차 비율은 1이다" 라는 영가설을 테스트해보자.
- "두 변수의 표준편차 비율은 1보다 작다" 라는 영가설을 테스트해보자.
- "두 변수의 표준편차 비율은 1보다 크다" 라는 영가설을 테스트해보자.
- 쌍체표본(paired sample)도 있긴 하지만 여기서는 더이상 다루지 않고 넘어간다.

결과표를 꼼꼼히 들여다보자.

- 비(ratio)는 어떻게 계산되었는가?
- F 값과 자유도는 어떻게 계산되었는가?
- 대립가설 별로 유의확률(p-value)은 어떻게 계산되었는가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석

일원분산분석

일원분산분석(one-way Analysis of Variance)은 여러 모집단의 분산에 대한 가설검정 기법이다.

- 사실 지금까지 설명한 분산의 추정과 비교는 일원분산분석을 위한 준비에 불과하다.
- 실용적 맥락에서 보면 일원분산분석은 여러 개의 표본들이 주어졌을 때 이것들이 하나의 모집단에서 추출되었는가를 확인하는데 쓰인다.
- 한편 분산분석(Analysis of Variance 또는 ANOVA)은 특히 실험적 연구의 맥락에서 분석을 위해 자주 사용되는 고도로 발달한 여러 분석기법들의 한 집합이다. 심리학 (그리고 경영학 일부)에서는 굉장히 중요한 기법들로 심도있게 공부된다.
- 일원분산분석은 이 분야의 여러 기법들의 가장 기초가 되는 부분이다. 기본적인 개념을 학습하는 것에 더하여 여러가지 맥락에서도 자주 사용된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석

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

$$\sum (Y - \bar{Y})^2 = \sum (Y - \bar{Y}_G)^2 + \sum (\bar{Y}_G - \bar{Y})^2$$

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

$$\sum \frac{(Y - \bar{Y})^2}{(n_{obs} - 1)}) = \sum \frac{(Y - \bar{Y}_G)^2}{n_{obs} - n_{group}} + \sum \frac{(\bar{Y}_G - \bar{Y})^2}{n_{group} - 1}$$

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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석

일원분산분석의 검정통계량(test statistic)은 F값이다.

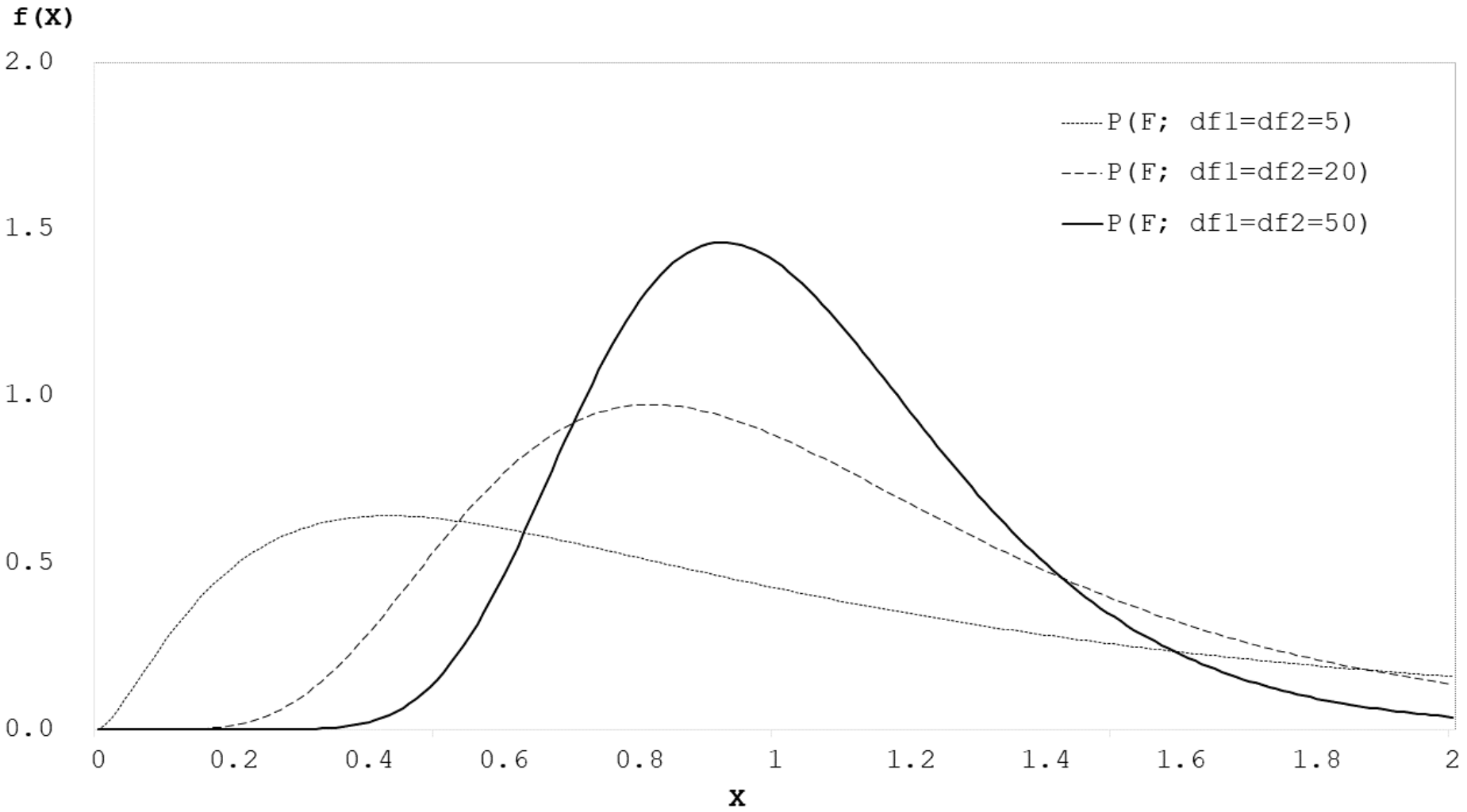
- 일원분산분석은 F 분포를 사용하며 그 원리는 앞서 설명한 두 모집단 분산에 대한 가설검정과 같다.

$$F_{(n_{group}-1, n_{obs}-n_{group})} = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/(n_{group}-1)}{SS_{within}/(n_{obs}-n_{group})}$$

- 분자인 $MS_{between}$는 실험처리(treatment)에 의한 효과로 볼 수 있고, 분모인 MS_{within}는 무선오차(random error)에 지나지 않는다.
- 왜 분모가 겨우 무선오차에 불과한가를 물을 수 있는데, 이는 일원분산분석이 모집단에 대해 정규성(normality)을, 이로부터 표본을 뽑을 때는 임의성(randomness)을, 마지막으로 뽑힌 표본들은 등분산성(homogeneity of variance)을 가정하기 때문이다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

일원분산분석



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석

일원분산분석의 가설설정은 다소 주의를 요한다.

- 일원분산분석의 영가설은 "모든 집단에 걸쳐 평균값이 동일하다"이다. 구체적인 예를 들면, "기후위험의 인식 점수의 평균에 최종학력별 차이는 없다"가 될 수 있다.
- 그 대립가설은 "모든 집단에 있어 평균값이 다르다"가 아니다. "적어도 하나의 집단에서 평균값이 다르다"임에 주의할 것!
- 이 영가설을 넘어 보다 구체적으로 둘 중 어느 집단에서 평균값이 더 높은가를 짝지어 비교하는 기법들도 있다. 예컨대 본페로니(Bonferroni), 셰퍼(Scheffe), 시닥(Sidak) 다중비교검정(multiple comparison test)이 있다. 우리는 다루지 않는다.
- 분석상 분산의 비율을 비교하고 있음에도 불구하고 가설설정이 평균에 대해 이루어진다는 점도 기억할 것.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

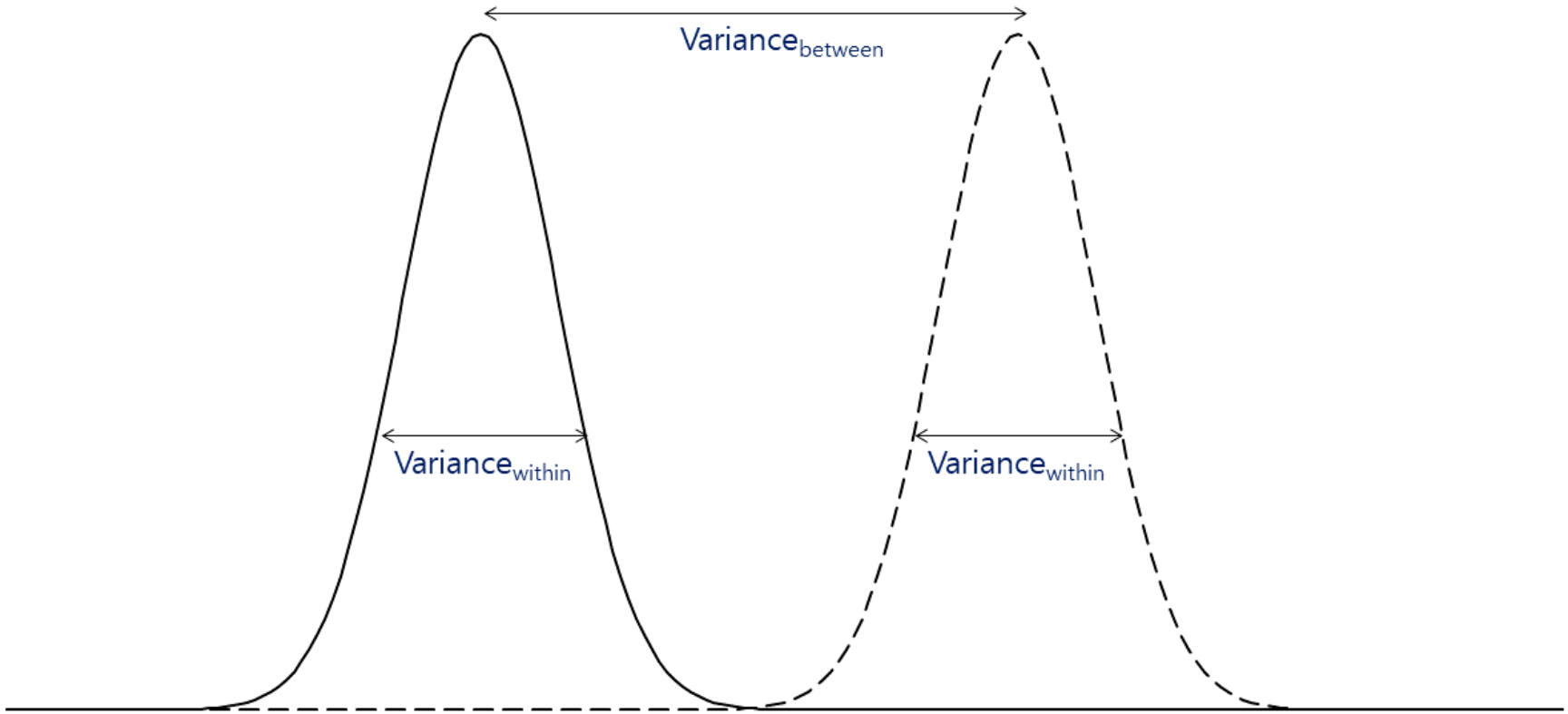
일원분산분석

계산된 F값과 두 개의 자유도에 따라 유의성 검정을 수행한다.

- 만일 모든 집단에 있어 평균값이 동일하다면, 즉 영가설대로라면 F값은 1이다(Why?).
- 영가설의 전제 아래 F 분포를 따르는 (가상적인) 표집분포를 그린 뒤 표본에서 뽑힌 F 값을 확인해보자. 만일 그 검정통계량이 1보다 매우 커서 오른쪽 끄트머리의 유의확률(p-value)이 0.05보다 작으면 95% 신뢰수준에서 영가설을 기각한다.
- 집단 간(between) 분산이 크고 집단 내(within) 분산이 작으면 F값이 커져서 영가설이 기각하는 경향이 있음을 확인할 수 있다.

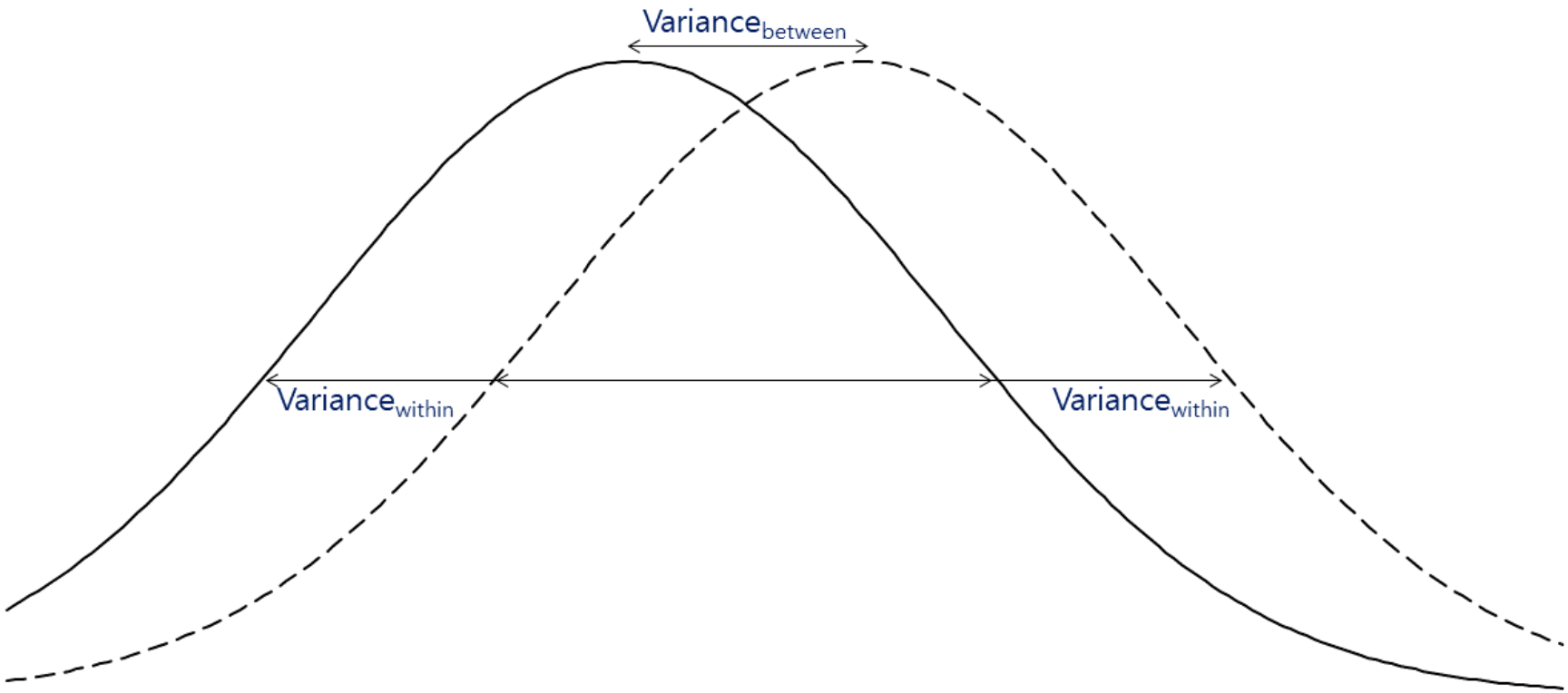
충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

일원분산분석의 실제 활용

경험적 연구논문에서 일원분산분석은 크게 두 부분에서 주로 활용된다.

- 첫번째는 표본에 관한 기술통계(descriptive statistics)를 제시하는 부분이고, 두번째는 회귀분석(regression analysis)에서 전체 계수(coefficients)의 유의성 검정(significance test) 부분이다.
- 먼저 연구자는 자신의 표본 안의 "핵심이 되는 범주형 관심변수 내지 종속변수에 따라" 다른 여러 변수들이 어떻게 달라지는지 일원분산분석을 통해 살펴볼 수 있다. 그 비율은 기술통계(descriptive statistics)의 일부로 보고할 수 있다. 이때 범주형 변수는 명목형 내지 순서형 척도로 측정된 것이며 예컨대 최종학력, 출신지역, 지지하는 정당 등을 생각해 볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

첫번째 맥락으로 활용된 한 논문의 기술통계 파트에서 일원분산분석이 실제로 어떻게 활용되는지 살펴보자

- 이희정 (2018)의 〈표7〉를 꼼꼼히 살펴보자. 이 표는 공정성 인식(perception toward justice)이 여러가지 청년의 인구학적 · 사회적 속성에 따라 어떻게 다른가를 보여준다.
- 공정성 인식에는 고용상태(employment)에 따른 차이가 있나? 그 차이는 95% 신뢰수준에서 통계적으로 유의한가?
- 공정성 인식에는 주관적 사회경제 지위(subjective socioeconomic status)에 따른 차이가 있나? 그 차이는 95% 신뢰수준에서 통계적으로 유의한가?

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

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

일원분산분석은 두번째 맥락인 회귀분석 전체 계수의 유의성 검정에서도 사용된다([Stata 코드] 참고).

- 이 경우 영가설은 "모든 회귀계수들이 0이다"로 만일 이 영가설을 기각하지 못한다면 모델에 포함된 어떠한 독립변수(X)로도 종속변수(y)를 의미있게 설명하지 못함을 의미한다. 당연히 이 경우에는 모델을 처음부터 다시 만들어야 한다.
- 회귀분석의 맥락에서 대립가설은 "적어도 하나 이상의 회귀계수는 0이 아니다" 임에 주의할 것.
- 이 맥락에서 영가설은 "회귀계수가 0이다($b = 0$)"로, 다시 말해 해당 독립변수는 종속변수를 설명하는데 의미가 없다는 뜻이다. 물론 연구자는 이 영가설을 기각하고 싶기 마련($b \neq 0$)이다. 더 자세한 내용은 몇 주 뒤에 다루게 된다.
- 이에 관한 더 자세한 내용은 몇 주 뒤에 다루게 된다. 하지만 정말로 사용된다는 점을 잠깐 확인해보기로 하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

t 검정은 두 모집단의 평균을 비교하고, 일원분산분석은 여러 모집단에 걸친 분산의 비율을 비교한다.(**[Stata 코드]** 참고).

- 만일 집단이 두 개만 주어졌을 때 일원분산분석을 수행하면 어떤 결과를 가져올까? 이 경우 일원분산분석의 영가설은 "모든 집단에 걸쳐 평균값이 동일하다" 였으므로 이는 다시 "두 집단에 걸쳐 평균값이 동일하다" 로 축소된다. 이것은 t 검정과 같은 것이 아닌가?!
- 실제로 F 값과 t 값에는 밀접한 관계가 있다.

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

- 이 사실만 간단히 Stata에서 확인해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

반대로 생각해서, 집단이 여러 개 있을 때 구태여 일원분산분석을 배우는 대신 t 검정을 여러번 하면 안될까?(**[Stata 코드]** 참고).

- 결론만 말하자면 (1) 굉장히 불편하고 혼란스러울 뿐 아니라, (2) 추정상의 오류를 저지르게 될 위험이 극단적으로 커지므로 권할 수 없다.
- 먼저 t 테스트를 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?).
- 게다가 이 10번의 t검정을 수행하는 과정에서 최소 1번 이상 오류가 나타날 가능성은 급격히 증가한다. 예컨대 5% 유의확률이라면 1회 이상의 오류 확률은 40%이나 된다(Why?).
- 이 사실을 간단히 Stata에서 확인해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

일원분산분석의 실제 활용

분산분석 계열의 기법은 사실 굉장히 다양한 테크닉들을 광범위하게 포괄한다.

- 일원분산분석에서 일원(one-way)는 독립변수가 하나인 것을 의미한다. 독립변수가 두 개인 경우 이원분산분석(two-way ANOVA)이, 그보다 많은 경우 다원분산분석(multi-way ANOVA)이 적절하다.
- 다변량분산분석(multivariate ANOVA 또는 MANOVA)은 종속변수가 여러 개인 경우에 사용된다.
- 그 밖에도 공분산분석(analysis of covariance 또는 ANCOVA)이나 다변량공분산분석(multivariate analysis of covariance 또는 MANCOVA) 등 다양하게 있으나 이것들은 이 수업의 수준을 벗어난 것이므로 더이상 다루지 않는다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY