사회통계

분산에 관한 가설검정

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. χ^2 분포와 자유도
2. 단일 모분산에 관한 가설검정
3. F 분포와 자유도
4. 두 모집단의 분산에 관한 가설검정

χ^2 분포와 자유도

모분산 σ^2에 관한 가설검정은 여러 상황에서 쓰일 수 있다.

1. 가령 소득의 분산은 소득 불균등을 의미한다(사실 사회계층 및 불평등 연구에서는 좀 더 정교한 소득 불균등 지표가 사용되기는 한다).
2. 한 반에서 학생 성적의 분산은 교육 성과가 고르게 나타나지 않아 교육 양극화가 나타남을 시사한다.
3. 설문조사의 특정 문항에 대한 응답의 분산은 조사대상자 의견의 불일치 내지 다양성을 의미한다.
4. 금융상품의 수익률의 분산은 해당 상품의 스프레드(spread) 또는 리스크(risk)를 의미한다. 똑같은 논리가 도박에 대해서도 적용된다.
5. 제조된 공산품의 상태지표의 분산은 들쭉날쭉한 품질(quality) 상태를 의미한다.

χ^2 분포와 자유도

일단 하나의 표본분산 s^2에 대해 이론적으로 상상해보자.

- 우리는 모집단에서 표본을 무한히 계속 뽑고 그것들의 분산 s^2를 구해 분산의 표집분포를 상상해 볼 수야 있다
- 이때 표본 크기 n이 작다면 분산의 표집분포는 오른쪽 꼬리가 길어진다(long tail). 그 이유는 생각보다 단순한데, 분산의 계산 과정에서 제곱을 하다보면 큰 값이 우연히 많이 나오기 때문이다(Why?).
- 하지만 아쉽게도 표본분산 s^2에 관한 직접적인 이론 분포는 없다!

χ^2 분포와 자유도

- 표본평균 $\bar{x}$의 경우 소표본일 때 t 분포, 대표본일 때 Z 분포라는 이론적 확률분포를 가졌다.
- 반면 분산 s^2의 경우 직접적인 이론적 확률분포가 없고, "표본분산 s^2와 모분산 σ^2 간의 비율을 자유도와 곱한 값"에 관한 이론적 확률분포인 χ^2 분포(chi-square distribution)를 대신 사용한다(교과서 175–177).

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

- 다시 한 번 강조하자면, 단일표본의 분산 s^2의 표집분포가 아니라, $(n-1)\frac{s^2}{\sigma^2}$의 표집분포를 상상해야 한다!

χ^2 분포와 자유도

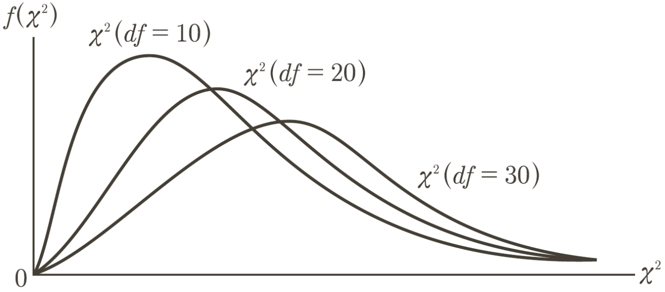
χ^2 분포는 오로지 자유도 $n-1$에 의해 모양이 결정된다.

- χ^2 분포는 소표본일 때 비대칭적이고 오른쪽으로 꼬리가 길다.
- 제곱을 하다보면 우연히 큰 값이 나타나기 쉽고 표본이 작다보면 그 값이 튀어보이기 때문이다(Why?).
- 그러나 대표본이 될수록 정규분포의 모습에 점점 근사한다.
- χ^2 값은 결국 분산들의 비율이므로 항상 양수값을 갖는다(Why?).

$$(n-1)\frac{s^2}{\sigma^2} \geq 0$$

- 만약 s^2가 σ^2보다 많이 크다면, χ^2는 1보다 확실히 커지고 검정통계량은 우측 끄트머리에 위치하게 된다!

χ^2 분포와 자유도



단일 모분산에 관한 가설검정

단일 모분산에 관한 χ^2 검정의 가설 구조를 살펴보자.

- 양측검정에서 귀무가설과 대립가설의 구조는 다음과 같다.

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_a : \sigma^2 \neq \sigma_0^2$$

- 단측검정은 두 가지 형태 중 하나의 귀무가설과 대립가설의 구조를 갖는다.

$$H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_a : \sigma^2 < \sigma_0^2$$

또는

$$H_0 : \sigma^2 \leq \sigma_0^2 \qquad H_a : \sigma^2 > \sigma_0^2$$

단일 모분산에 관한 가설검정

앞서 배운 바와 마찬가지로 유의성 검정을 수행할 수 있다.

- 가장 먼저 모분산 σ^2에 대한 귀무가설 및 대립가설을 설정한다(교과서 195–198).
- 자유도 df, 표본분산 s^2, (귀무가설로 설정된) 모분산 σ^2에 따라 다음과 같이 χ^2 값을 계산한다.

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

- 표본분산 s^2가 모분산 σ^2보다 훨씬 클수록 χ^2 분포 위에서 극단적인 값을 갖게 되므로 유의확률(p-value)은 작아진다(Why?).
- χ^2 분포의 "곡선 밑 면적" 색칠공부는 엑셀 함수 CHISQ.DIST(x, deg_freedom, cumulative)를 사용한다.
- 주어진 유의수준에 따라 귀무가설을 기각할 수 있는지 확인하고 결과를 보고한다.

단일 모분산에 관한 가설검정

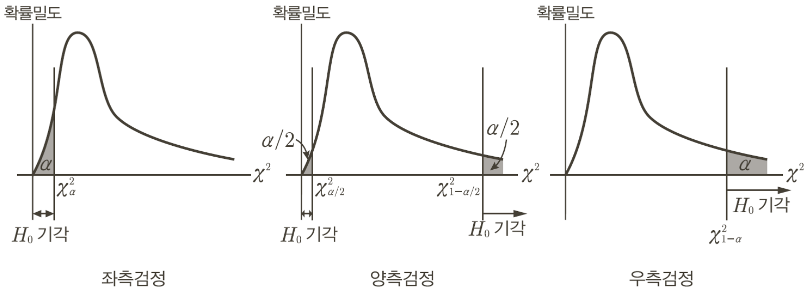
예제 1. 새우는 50년 전통과 까다로운 품질관리를 자랑하는 통조림 회사에 근무하고 있다. 최근 그는 생산2팀으로부터 자사 통조림 용량이 정규분포하고 그 분산은 16이라는 보고를 받았다. 새우는 생산2팀 주장의 진위를 확인하기 위해 28개의 표본을 임의로 선정하였다. 조사 결과 그 분산은 25임을 발견하였다. 이 소식을 전해들은 생산2팀장 고래는 노발대발하며 "야 이놈아, 그건 니 표본이 이상해서 그런거고! 원래 우리 팀에서 만드는 전체 통조림의 분산은 16이 맞아!" 발끈한 새우는 그 말이 사실인지 끝까지 검정하고 싶어졌다. 새우를 위한 가설을 제시하고 95% 신뢰수준에서 이를 검정하시오.

단일 모분산에 관한 가설검정

- 가설이 어떻게 세워져야 할지 두 가지로 생각해보자. 먼저 첫번째는 아래와 같다.

$$H_0 : \sigma^2 \leq 16$$
$$H_a : \sigma^2 > 16$$

- χ^2 분포 위에 색칠공부도 해보자. 어느 부분에 검정통계량인 χ^2 값이 놓여야 할까?



단일 모분산에 관한 가설검정

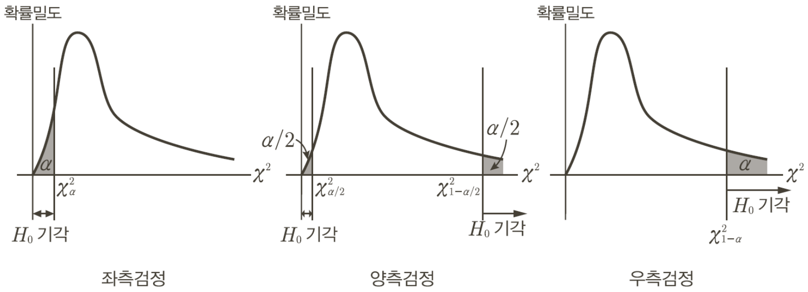
- χ^2 값은 $\dfrac{(n-1)s^2}{\sigma^2} = \dfrac{(28-1)\cdot 25}{16} = 42.1875$이다.
- 자유도는 27 (=28 – 1)이고 단측검정이므로 엑셀에서 1-CHISQ.DIST(42.1875, 27, TRUE)로 유의확률(p-value)을 계산한다(Why?).
- 이 유의확률은 귀무가설에 따라 그려진 χ^2 분포에서 검정통계량보다 극단적인 값을 얻게 될 확률이다. 이것이 아주 작다면 귀무가설을 '자신있게' 기각할 수 있다.
- 새우가 얻는 유의확률은 0.0315 정도로 0.05보다 작다. 그러므로 모집단의 분산이 16이라는 귀무가설은 95% 신뢰수준에서 기각된다.
- 즉 생산2팀에서 만들어진 통조림 용량의 분산은 16보다 통계적으로 유의하게 크다.

단일 모분산에 관한 가설검정

- 다른 방식으로 가설을 세워보자.

$$H_0 : \sigma^2 = 16$$
$$H_a : \sigma^2 \neq 16$$

- χ^2 분포 위에 색칠공부도 해보자. 어느 부분에 검정통계량인 χ^2 값이 놓여야 할까?



단일 모분산에 관한 가설검정

- 이때는 유의확률(p-value)의 계산이 원활하지 않다(Why?).
- 따라서 95% 신뢰구간을 직접 구해보자. 엑셀에서 CHISQ.INV(0.025,27)를 통해 신뢰하한을, =CHISQ.INV(0.975,27)를 통해 신뢰상한을 구할 수 있다.
- 이때 신뢰하한과 신뢰상한의 절대값은 같지 않음에 주의하자(Why?).
- 만일 여러분이 구한 검정통계량이 이 구간 밖에 놓인다면, 귀무가설을 통계적으로 유의하게 기각할 수 있다.
- 그러나 검정통계량은 신뢰구간 안에 있다. 그러므로 양측검정 결과에 따르면 생산2팀에서 만들어진 통조림 용량의 분산은 16과 통계적으로 유의하게 다르다고 할 수 없다.

예제 2. 고래는 시멘트를 생산하는 공장에서 통계적 품질 관리자로 근무하고 있다. 품질분임조(QC)에서 결정된 사항은 시멘트 한 포대의 무게는 평균적으로 40kg, 표준편차도 0.5kg 이하를 유지하는 것이다. “혹시 내 생산라인의 표준편차가 저것보다 큰 게 아닐까?” 여러모로 불안해진 고래는 헐레벌떡 자기 생산라인으로 뛰어와 10포대의 시멘트를 랜덤하게 추출하여 무게를 재어본 뒤, 그 자료를 cement.csv라는 파일로 업로드하였다. 과연 고래의 생산라인은 품질관리를 똑바로 하고 있는지 1% 유의수준에서 검정하시오.

단일 모분산에 관한 가설검정

- 가설은 다음과 같다(Why?).

$$H_0 : \sigma^2 \leq 0.25$$
$$H_a : \sigma^2 > 0.25$$

- χ^2 값은 다음과 같다.

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(10-1)\cdot 0.669}{0.25} = 24.084$$

- 직접 χ^2 분포를 그리고 색칠공부 영역을 확인해보자. χ^2 값도 그 위치에 표시하자.

단일 모분산에 관한 가설검정

- 자유도는 9 (=10 – 1)이고 단측검정이므로 엑셀에서 1-CHISQ.DIST(24.084, 9, TRUE)로 유의확률을 계산한다.
- 유의확률은 0.00417 정도로 0.01보다 작다. 그러므로 모집단의 분산이 0.25라는 귀무가설은 1% 유의수준에서 기각된다.
- 다시 말해, "고래의 생산라인에서 만들어진 시멘트 한 포대의 무게 표준편차는 99% 신뢰수준에서 통계적으로 유의하게 0.5보다 크다."

단일 모분산에 관한 가설검정

사실 χ^2 분포는 중요한 가정을 전제한다.

- 첫째, (수학적으로 세밀하게 짚고 넘어가지는 않지만) χ^2 분포가 성립하기 위해서는 모집단이 정규분포해야 한다(Why?).
- 만일 모집단의 정규성 가정(normality assumption)이 위배되는 상황이라면 함부로 χ^2 분포를 사용한 통계적 추론을 수행할 수 없다.
- 실무나 연구 상황에서는 (1) 이리저리 조사하여 모집단이 정규분포하는지 확인해 볼 수 있을 것이다.
- 모집단에 관한 단서가 전혀 없다면, (2) 최대한 표본의 크기를 늘려($n > 100$ 이상) 문제를 줄인다.
- 만일 모집단에 관한 단서를 전혀 확보할 수 없고 표본 크기도 늘릴 수 없는 상황이라면, (3) 주어진 임의표본을 가지고 분위수 대조도(quantile-quantile plot; QQ plot) 같은 정규성 검정(normality test)을 수행한다.

단일 모분산에 관한 가설검정

- 둘째, 표본(=변수)들은 서로 독립적으로 추출되어야 한다.
- 이 가정없이는 다음과 같은 χ^2 값의 계산 자체가 불가능해진다(Why?).

$$\chi_k^2 = Z_1^2 + Z_2^2 + ... + Z_k^2 = \sum \left(\frac{x-\mu}{\sigma}\right)^2$$

- 이 가정은 다음 주 카이제곱 검정(chi-square test)에서 매우 유용하게 사용된다.

F 분포와 자유도

F 분포와 자유도

모집단의 분산에 관해서도 두 가지로 테스트해 볼 수 있다.

- 만약 모집단이 “하나” 있다고 하자. 이것의 분산의 표집분포가 χ^2 분포를 따른다고 할 때, 귀무가설을 나의 표본에 따라 기각할 수 있는가 살펴볼 수 있다. 이것이 (아까 배운) 단일표본 χ^2 검정(one-sample chi-square test)이다.
- 만약 모집단이 “두 개” 주어졌다고 하자. 두 모집단 분산의 비율(ratio)의 (가상적인) 표집분포가 F 분포(F distribution)를 따른다고 할 때, 귀무가설을 나의 표본에 따라 기각할 수 있는가 살펴볼 수 있다. 이것을 F 검정(F test)이라고 부른다.
- 말하자면 평균비교(mean comparison)처럼 분산비교(variance comparison)도 수행할 수 있는 셈이다!

F 분포와 자유도

분산비교는 F 분포를 이용한다.

- "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(probability density function; PDF)는 다음과 같이 정의된다.

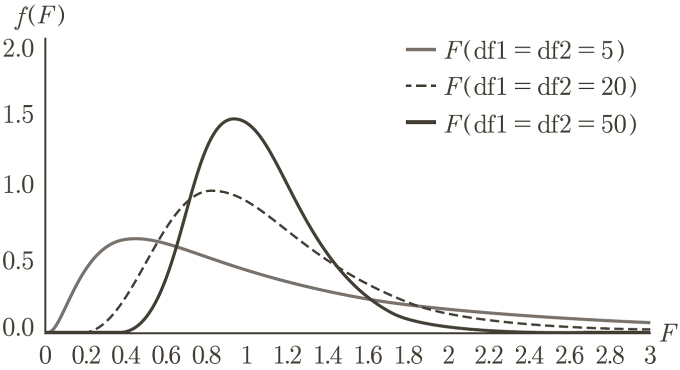
$$F_{(n_1-1,n_2-1)} = \frac{\chi_1^2/(n_1-1)}{\chi_2^2/(n_2-1)}$$

- 조금만 생각해보면 F 분포는 각각의 "단일모집단의 분산"을 설명하는 χ^2 의 비율임을 쉽게 파악할 수 있다.

F 분포와 자유도

- 우리는 두 표본의 분산의 차(difference)를 구하려는 것이 아니라 비(ratio)를 구한다. 따라서 두 χ^2 값이 서로 비슷할수록 F 값은 (0이 아니라) 1에 접근한다.
- F 값은 두 개의 χ^2 분포의 비로 구성되므로 항상 양수(+)이다.
- F 분포는 두 개의 자유도 $(n_1 - 1)$과 $(n_2 - 1)$을 패러미터(parameter)로 받아 그 형태가 결정된다(이것도 χ^2가 하나의 자유도 $n - 1$를 패러미터로 받아 그 형태가 결정되는 것과 유사하다).
- 두 개의 자유도가 작을 때는 우측 꼬리가 길다. 두 개의 자유도가 점점 커지면 정규분포에 점점 근사한다.
- χ^2 분포와 마찬가지로 대칭적인 형태가 아니다.
- 일단 χ^2 분포 자체에 익숙해져야 F 분포에 대해서도 쉽게 이해할 수 있다.

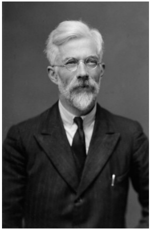
F 분포와 자유도


$f(F)$
2.0
1.5
1.0
0.5
0.0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 2.4 2.6 2.8 3
F
$F(\text{df1} = \text{df2} = 5)$
$F(\text{df1} = \text{df2} = 20)$
$F(\text{df1} = \text{df2} = 50)$

F 분포와 자유도

F 분포는 Ronald Fisher의 이름을 따서 만들어졌다.

- Fisher는 천재였다. (약간의 과장들을 덧붙인다면) 6세에 수학에 몰두하기 시작하여 22살에는 최대우도법(Maximum Likelihood Estimation; MLE)의 수학적 원리에 기여하기도 했다. 우리가 배운 유의확률(p-value), F 분포, 가설검정의 논리, (곧 배울) 분산분석(analysis of variance; ANOVA) 등등 현대통계학의 초석을 쌓았다.
- 성격은 좀 그저 그랬던 듯 하다. 평생에 걸쳐 또다른 천재 Karl Pearson과 불화를 겪었다.



두 모집단의 분산에 관한 가설검정

검정통계량 F 값을 계산할 때 몇 가지 주의할 점이 있다.

- 우리가 추정하고자 하는 것은 두 모집단의 분산 비율 σ_1^2/σ_2^2 임을 기억해야 한다.
- 정규분포하는 두 모집단에서 서로 독립인 임의표본을 추출한 뒤, 모분산(population variance)의 불편추정량 s_1^2과 s_2^2를 다음과 각각 같이 계산한다.

$$s_1^2 = \frac{\sum(x_{1i} - \bar{x}_1)^2}{n_1 - 1}, \qquad s_2^2 = \frac{\sum(x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

- 그러면 자연스럽게 F 값은 다음과 같이 단순화된다(Why?).

$$F_{(n_1-1, n_2-1)} = \frac{s_1^2}{s_2^2}$$

두 모집단의 분산에 관한 가설검정

- 검정통계량 F 값은 비율이기 때문에 어느 쪽이 분자가 되는지 주의해야 한다.
- 많은 교과서에서 분산이 큰 쪽을 반드시 분자로 놓도록 가르친다. 그러면 F 값은 반드시 1 또는 그보다 커지므로 양측검정은 사라지고 오로지 단측검정만 생각하면 된다(Why?).
- 우리도 이 설명 방식을 따른다!
- 그러므로 분자분모 설정에 따라 분산비교를 위한 F 검정의 가설구조는 오로지 우측 단측검정만을 따르게 된다.

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq 1$$

$$H_a : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

두 모집단의 분산에 관한 가설검정

F 검정의 기본적인 가정들에 대해서도 기억해 둘 필요가 있다.

- (1) F 분포는 (χ^2 분포를 이용하므로) 당연히 정규성 가정이 요구된다.
- 필요에 따라 모집단 분포에 관해 좀 더 조사하거나, 표본 크기를 늘리거나, 표본에 대해 정규성 검정을 수행하여 이 가정을 타당화한다.
- (2) F 분포는 (χ^2 분포를 이용하므로) 당연히 상호독립성 가정, 즉 두 표본이 독립적으로 뽑혔다고 전제된다.
- 이 가정이 위배되면 좀 더 특수한 분석을 수행하여야 하나 그것은 우리 수업의 범위를 벗어나므로 다루지 않는다.

두 모집단의 분산에 관한 가설검정

F 검정에서도 마찬가지로 유의성 검정을 수행할 수 있다.

- 가장 먼저 모분산 비율 σ_1^2/σ_2^2에 대한 귀무가설 및 대립가설을 설정한다. 분자분모를 주의깊게 고르면 가설구조는 늘 똑같다.
- (정규분포한다고 가정된) 모집단으로부터 작은 표본을 뽑는다. 자유도 두 개 df_1, df_2, 표본분산 비율 S_1^2/S_2^2, (귀무가설로 설정된) 모분산 비율 σ_1^2/σ_2^2에 따라 F 값, 즉 검정통계량을 계산한다.
- 이 검정통계량이 (자유도 두 개에 따라 형태가 결정된) F 분포 위 어디에 위치해 있는가에 따라 유의확률(p-value)을 계산할 수 있다.
- 주어진 유의수준에 따라 귀무가설을 기각할 수 있는지 확인하고 결과를 보고한다.
- F 분포의 "곡선 밑 면적"을 계산하는 엑셀 함수는 F.DIST(x, deg_freedom1, deg_freedom2, TRUE)이다.

두 모집단의 분산에 관한 가설검정

예제 3. 아동교육시스템에 관해 연구하는 새우는 부산 소재 어린이집 명단인 daycare.csv를 들여다보다 우연히 평가 인증 여부에 따라 아동 수의 분산이 상이하다는 생각하고 있다. "평가인증을 받은 어린이집에서는 (그렇지 않은 어린이집보다) 아동 수의 분산이 큰 것 같아." 아동 수의 모집단이 정규분포한다는 가정 아래, 새우를 위한 귀무가설과 대립가설을 제시하고 1% 유의수준에서 이를 검정하시오.

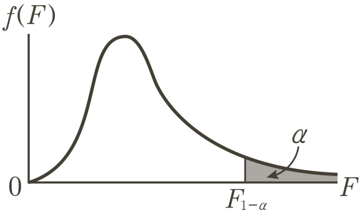
두 모집단의 분산에 관한 가설검정

- 가설은 다음과 같다(Why?).

$$H_0 : \frac{\sigma_Y^2}{\sigma_N^2} \leq 1$$

$$H_a : \frac{\sigma_Y^2}{\sigma_N^2} > 1$$

- F 분포를 그리고 색칠공부를 해보자. 어느 부분에 검정통계량인 F 값이 놓여야 할까?



두 모집단의 분산에 관한 가설검정

- 자료를 먼저 평가 인증 여부에 따라 정렬(sort)한다. 그 다음 [데이터]-[데이터 분석]을 통해 "기술 통계량"을 평가인증 여부 별로 확인해보자.
- F 값을 계산해보면 $\dfrac{S_Y^2}{S_N^2} = \dfrac{1732.93}{846.13} = 2.05$ 이다.
- 두 자유도는 각각 121, 83 이므로(혼동 주의!), 엑셀에서 1-F.DIST(2.05, 121, 83, TRUE)로 유의확률(p-value)을 계산한다.
- 새우가 얻은 p-value는 0.0003 정도로 0.01보다 다소 작다. 그러므로 귀무가설을 1% 유의수준에서 기각할 수 있다.
- 다시 말해, 평가 인증을 받은 어린이집은 (그렇지 않은 어린이집보다) 아동 수의 분산이 통계적으로 유의하게 크다.

두 모집단의 분산에 관한 가설검정

- 이번에는 데이터 분석 기능을 활용해 좀 더 쉽게 해보자.
- 엑셀에서 [데이터]-[데이터 분석]을 통해 "F-검정: 분산에 대한 두 집단"을 선택하자.
- 여기서 주의할 부분은 변수 1이 무조건 분자, 변수 2가 무조건 분모로 들어간다는 사실이다. 다시 말해, 분산이 큰 쪽을 자동적으로 분자에 넣어주거나 하지 않는다는 점이다.
- 이런건 억지로 외울 필요가 없다. 일단 해보고 만일 F 값이 1보다 작게 나왔다면 "아 바뀌었군" 하면서 다시 하면 된다.
- 데이터 분석 결과, $F = 2.05$ 그리고 $p = 0.0003$ 임을 재확인할 수 있다.