사회통계
평균비교

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 두 모평균에 관한 가설검정
2. 쌍체표본 t 검정
3. 독립표본 t 검정
4. 등분산 가정
5. 두 모비율에 관한 가설검정

두 모평균에 관한 가설검정

두 모평균에 관한 가설검정

보통 평균비교에 관한 가설검정은 t 검정으로 수행한다.

- 단일표본 t 검정을 통해 소표본으로 "하나의 모집단의 평균 μ이 어떤 특정 값 μ_0인가"를 살펴볼 수 있었다.
- 표본 크기가 충분히 크면($n \geq 30$) 평범하게 표준정규분포(Z 분포)를 사용해서 가설검정할 수 있다.
- 표본 크기가 작으면($n < 30$) t 분포를 사용해서 가설검정할 수 있다(단일표본 t 검정).
- 그런데 t 검정(t test)을 말할 때는 보통 두 변수가 주어져 있고 "두 모평균의 차이 $\mu_1 - \mu_2$가 0과 같은가/큰가/작은가"를 살펴보는 통계분석 기법을 의미한다(독립표본 t 검정과 쌍체표본 t 검정).

두 모평균에 관한 가설검정

평균비교는 두 모집단의 평균의 차이를 가설검정의 대상으로 삼는다.

- 평균비교는 양측검정과 단측검정으로 수행될 수 있다.

$$H_0 : \mu_1 - \mu_2 = 0 \qquad H_0 : \mu_1 - \mu_2 \geq 0 \qquad H_0 : \mu_1 - \mu_2 \leq 0$$
$$H_a : \mu_1 - \mu_2 \neq 0 \qquad H_a : \mu_1 - \mu_2 < 0 \qquad H_a : \mu_1 - \mu_2 > 0$$

- 사실 우변에서 0이 아닌 숫자로 가설검정을 할 수 있다. 하지만 실무나 연구에서 0이 많이 쓰이므로 이걸로 충분하다.

두 모평균에 관한 가설검정

인문사회·공학·보건의료 등 여러 분야에서 평균비교는 매우 유용하다.

- 남성의 연 평균소득(μ_1)과 여성의 연 평균소득(μ_2)의 차이
- 전업 공연예술인의 예술지원정책 만족도(μ_1)과 비전업인의 정책 만족도(μ_2)의 차이
- 가출경험이 있는 중학생이 가진 비행경험 친구의 수(μ_1)와 가출경험이 없는 중학생이 가진 비행경험 친구의 수(μ_2)의 차이
- 집중강화 트레이닝을 받기 전 2군 선수들의 이전 평균기록(μ_1)과 이후 평균기록(μ_2)의 차이
- 처방약을 복용한 환자의 혈압(μ_1)과 위약을 복용한 환자의 혈압(μ_2)의 차이

두 모평균에 관한 가설검정

독립표본 t 검정과 쌍체표본 t 검정은 어떻게 다른가?

- 이것은 데이터가 어떻게 생겼는가로 쉽게 이해할 수 있다.
- 왼쪽은 쌍체표본(paired samples)으로 같은 사람에 대해 처방 전후(before and after)로 기록이 짝지어(paired) 있는 반면, 오른쪽은 독립표본(independent samples)은 처방(treatment) 실시 여부를 말해주는 가변수(dummy variable)가 있다.

ID	BEFORE	AFTER
1	35	27
2	31	39
3	46	33
4	39	40
5	31	31

ID	TREATED	RECORD
1	0	35
2	0	31
3	0	46
4	0	39
5	0	31
1	1	27
2	1	39
3	1	33
4	1	40
5	1	31

쌍체표본 t 검정

쌍체표본 t 검정

예제 1. social_paired.csv를 엑셀로 불러오시오. 아동 사회화에 관한 가족사회학 연구를 수행하는 고래는 초등학생의 사회적 자아가 1년 전후로 분명히 차이가 있을 것이라고 예상하고 있다. 각각의 시점에서 사회적 자아 점수는 socialself1과 socialself2 변수로 입력되어 있다. 고래의 귀무가설과 대립가설을 제시하고 이를 95% 신뢰수준에서 검정하시오.

쌍체표본 t 검정

- 고래의 적절한 가설은 다음과 같으며 양측검정이 필요하다.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

- 자료를 열어보면 표본 수가 매우 작고 쌍체표본이므로 쌍체표본 t 검정이 적합하다.
- 하나하나 엑셀 함수를 사용해 t 검정을 수학적으로 멋지게 풀 수 있다. 하지만 이제 신물이 난다. 다 때려치우고 [데이터]-[데이터 분석]을 따라가 데이터 분석 메뉴에서 "t-검정: 쌍체비교"을 선택하자.
- 변수 1과 변수 2의 입력 범위를 선택하고 (하이라이트 범위에 따라) "이름표" 체크에도 주의한다. "가설 평균차"에는 0을 입력한다(Why?). "유의 수준"은 0.05를 그대로 내버려 둔다.

쌍체표본 t 검정

- t 통계량으로 -2.132를 얻었다. 모집단 평균 차 $\mu_1 - \mu_2$가 0이라는 귀무가설에 따라 평균 차이의 표집분포를 그렸을 때, 이보다 더 극단적인 t 값을 얻을 확률, 즉 유의확률(p-value)은 얼마인가?
- 이것은 양측검정이므로 T.DIST(-2.132,14,TRUE)와 1-T.DIST(2.132,14,TRUE)를 더해야 한다. 이제 유의확률(p-value)이 대략 0.051임을 알 수 있다.
- 구해진 유의확률이 0.05보다 약간 크므로 고래는 자신의 귀무가설을 95% 신뢰수준에서 기각할 수 없다.
- 결론적으로 초등학생의 사회적 자아는 1년 전후로 통계적으로 유의한 차이를 발견할 수 없었다.

쌍체표본 t 검정

- 만일 귀무가설이 $H_0 : \mu_1 - \mu_2 \leq 0$로 단측검정이라면 어땠을까?
- T.DIST(-0.482,14,TRUE)로 구할 수 있는 유의확률(p-value)은 약 0.319이다. 이렇게 해도 어쨌든 귀무가설을 기각하지는 못한다.
- 한편 엑셀에서는 단측검정과 관련하여 하나 밖에 결과를 주지 않았다. 그건 반대쪽 꼬트머리의 유의확률(p-value)을 구할 때 어차피 1에서 다른 한쪽의 p-value를 빼면 되기 때문이다(Why?).
- 힌트는 t 분포도 정규분포와 마찬가지로 대칭적이라는 점이다.

쌍체표본 t 검정

쌍체표본 t 검정에서 검정통계량은 대체 어떻게 계산되는가?

- 위 예제 1에서는 엑셀 데이터분석을 사용하였을 뿐, 계산 원리는 다루지 않았다. 계산 원리를 잠깐 살펴보자.
- 앞서 단일표본 t 검정에서는 t 값을 다음과 같이 표준화하였다.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- 두 모집단의 평균 차이를 쌍체표본에서 비교할 때는 t 값을 이렇게 표준화한다.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_D/\sqrt{n}}$$

- 그런데 귀무가설로 $H_0 : \mu_1 - \mu_2 = 0$이라고 했으니 분자의 뒷부분은 없는거나 마찬가지다!
- 분모의 s_D는 표본평균 차이의 표준편차를 뜻한다.

독립표본 t 검정

독립표본 t 검정

예제 2. social_independent.csv 파일을 엑셀로 불러오시오. 고래와 같은 수업을 듣는 새우도 초등학생의 사회적 자아가 1년 전후로 분명히 차이가 있을 것이라고 귀무가설을 세웠다. 새우가 준비한 이 자료를 사용하여 위 가설을 95% 신뢰수준에서 검정하시오.

독립표본 t 검정

- 새우의 적절한 가설은 다음과 같으며 양측검정이 필요하다.

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

- 자료를 열어보면 표본 수가 매우 작고 독립표본이므로 독립표본 t 검정이 적합하다.
- 이번엔 [데이터]-[데이터 분석]을 거쳐, 데이터 분석 메뉴에서 "t-검정: 등분산 가정 두집단"을 선택하자.
- 변수 1과 변수 2의 입력 범위를 선택하고 (하이라이트 범위에 따라) "이름표" 체크에도 주의하자. "가설 평균차"에는 0을 입력한다(Why?). "유의 수준"은 디폴트로 내버려둔다(Why?).

독립표본 t 검정

- t 통계량으로 -1.816을 얻을 수 있다. 모집단 평균 차 $\mu_1 - \mu_2$가 0이라는 귀무가설에 따라 평균 차이의 표집분포를 그렸을 때, 이보다 더 극단적인 t 값이 나올 확률(p-value)은 얼마인가?
- 이것은 양측검정이므로 T.DIST(-1.816,28,TRUE)와 1-T.DIST(1.816,28,TRUE)를 더해야 한다. 이때 자유도는 2를 뺀다(Why?). 새우가 찾는 유의확률(p-value)은 대략 0.08임을 알 수 있다.
- 구해진 유의확률이 0.05보다 살짝 크므로 새우는 자신의 귀무가설을 95% 신뢰수준에서 기각할 수 없다.
- 결론적으로 1년 전후로 초등학생의 사회적 자아에서는 통계적으로 유의한 차이가 없었던 셈이다.

독립표본 t 검정

독립표본 t 검정에서 검정통계량은 대체 어떻게 계산되는가?

- 두 모집단의 평균 차이를 쌍체표본으로 비교할 때는 t 값을 이렇게 표준화하였다.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_D/\sqrt{n}}$$

- 두 모집단의 평균 차이를 독립표본으로 비교할 때는 t 값을 이렇게 표준화한다.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- 단일표본 t검정과 어떻게 다른지도 주의해서 살펴보자!

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

독립표본 t 검정

- 그런데 고래와 새우는 똑같은 자료를 가지고 가설을 검정하였는데 살짝 다른 검정통계량을 얻었다. 어떻게 이럴 수 있을까?
- 자료구조가 다르면 t 검정을 수행해도 t 값과 유의확률(p-value)이 달라진다(다시 확인해보자).
- 그 이유는 t 값을 계산할 때 (쌍체표본과 독립표본인가에 따라) 분모의 통합분산 (pooled variance)의 계산 방식이 다르기 때문이다.
- 이런 기술적 문제에 관해 시험이나 퀴즈로 낼 생각은 전혀 없다. 다만 의미가 무엇인지 곱씹어보자!

등분산 가정

등분산 가정

t 검정은 우선 두 표본의 분산이 같다고 가정한다.

- 결과표를 들여다 보면 타이틀에 "t-검정: 등분산 가정 두 집단"라고 적혀있다. 다시 말해, "두 표본이 나온 모집단들의 분산이 같다"고 전제된 것이다.
- 이를 등분산(homogeneity of variance) 가정 이라고 부른다.
- 만일 실험 전후로 같은 사람을 짝지워 평균을 비교하는 경우라면 이 가정은 (상대적으로) 크게 문제되지 않을 수 있다.
- 우리가 사용한 자료는 같은 초등학생을 1년 전후로 추적하여(before vs after) 사회적 자아를 조사한 것이므로 이 가정의 타당성이 별로 의심스럽지는 않다(Why?).
- 하지만 (쌍체표본 t 검정과는 달리) 독립표본 t 검정을 사용한다면 반드시 같은 사람끼리 짝지어 평균을 비교하지 않을 수도 있다.
- 다른 사람이라면 이 가정의 타당성이 특히 의문스럽고 이 가정은 아예 완화되어야 할 필요가 있다.

등분산 가정

예제 3. social.csv 파일을 다시 엑셀로 불러오시오. 젠더 수업을 듣는 참깨는 초등학생의 사회적 자아에 있어 성차가 있으리라는 귀무가설을 세웠다. 이 가설을 95% 신뢰수준에서 검정하시오.

등분산 가정

- 참깨의 적절한 가설은 다음과 같으며 양측검정이 필요하다.

$$H_0 : \mu_{\text{여}} - \mu_{\text{남}} = 0$$
$$H_a : \mu_{\text{여}} - \mu_{\text{남}} \neq 0$$

- 자료를 열어보면 표본 수가 매우 작고 독립표본이므로 독립표본 t 검정이 적합하다.
- (코드북을 보면 알 수 있듯) gender가 1이면 남자, 2이면 여자이고, (보통 사람의 성별이 1년 전후로 바뀌지 않으므로) 이것은 명백히 같은 사람을 짝지은 것은 아니다.
- 그래도 독립표본 t 검정에 의해 평균비교는 할 수 있다(쌍체표본 t 검정과는 이 점에서 크게 다르다). 다만 등분산 가정을 사용할 수 없음을 눈치채야 한다.

등분산 가정

- 이번에는 [데이터]-[데이터 분석]을 따라가 데이터 분석 메뉴에서 “t-검정: 이분산 가정 두집단”을 선택한다.
- 변수 1과 변수 2의 입력 범위를 선택하기 전에 gender가 너무 이리저리 섞여있음을 확인하자.
- 분석 메뉴를 끄고 자료로 되돌아와 [데이터]-[정렬]을 눌러 gender에 따라 정렬(sort)한다.
- 도로 [데이터]-[데이터 분석]을 선택하고 “t-검정: 이분산 가정 두집단”을 선택한다.
- 이제 변수 1과 변수 2의 입력 범위를 조심스럽게 선택하고 (하이라이트 범위에 따라) “이름표” 체크에도 주의하자. “가설 평균차”에 0을 입력하고 “유의 수준”은 디폴트로 내버려둔다.

등분산 가정

- 결과를 보면 해석이 t 값이 -0.28 정도 나왔다. 양측검정의 유의확률은 0.784이므로 너무 커서 5% 유의수준에서 귀무가설을 기각할 수 없다.
- 결론적으로 송이는 초등학생의 사회적 자아에 남녀 차가 없다는 귀무가설을 기각할 수 없었던 셈이다.
- 다만 실제 대규모 자료분석에 따르면 여아의 사회적 자아가 남아의 사회적 자아보다 99.9%의 신뢰수준에서도 통계적으로 유의하게 높다.

두 모비율에 관한 가설검정

두 모비율에 관한 가설검정

아까는 평균비교를 살펴보았고 이번엔 비율비교를 살펴보자.

- 앞서 살펴보았듯 평균비교는 두 모집단의 평균을 비교한다. 그 가설은 "두 모집단 간 평균의 차이는 없다" 또는 "두 모집단 간 평균의 차이는 0보다 크다/작다" 하는 식으로 설정된다.
- 반면 비율비교(proportion comparison)는 두 모집단의 비율을 비교한다.
- 예컨대 몇몇 지역에 걸친 환경재난 사건의 발생과 기후 문제가 가장 심각한 문제(Most Important Problem; MIP)라고 인식하는 사람의 비율 사이의 연관성을 살펴본다고 하자.
- 이 경우 "재난이 발생한 지역과 그렇지 않은 지역 사이에 기후 문제를 MIP라고 보는 사람의 비율은 차이가 없다"로 귀무가설이 설정된다(Why?).

두 모비율에 관한 가설검정

아까 평균비교를 했는데 왜 비율비교는 왜 해야하나?

- 기초 교과서 수준에서 설명하자면 평균 개념은 오로지 양적 변수에 대해서만 의미를 갖는다. 가령 양적 변수인 키나 몸무게의 평균은 의미를 갖지만, 질적 변수인 인종(1=백인; 2=흑인; ...)이나 종교(0=없음; 1=기독교; 2=불교; ...)의 평균에는 의미가 없다.
- 반면 질적 변수에서는 비율이 의미를 갖는다(e.g., 여성의 비율, 백인의 비율, 기독교의 비율 등).
- 수학적으로 볼 때, 평균비교는 정규분포나 t 분포를 사용하지만, 비율비교는 이항분포(binomial distribution)를 사용한다.
- 이에 따라 몇몇 교과서에서는 이항분포를 좀 더 깊이 다루면서 비율비교를 위한 공식을 제시하고 풀이과정을 연습한다.

두 모비율에 관한 가설검정

- 그런데 대규모 표본을 주로 분석하면 이항분포의 정규근사(normal approximation to the binomial)가 자연스럽게 활용될 수 있어 비율(이항분포)이 평균(정규분포)과 매우 비슷해진다.
- 소표본이라면 여전히 비율비교는 평균비교와 다른 개념이고 별도의 기법으로서 다루어져야 한다. 두 분석 결과가 제법 크게 달라지기 때문이다.
- 물론 여러분이 보건의료 분야에서 데이터 분석 전문가가 된다면 이야기는 달라진다.