사회통계

유의확률을 이용한 가설검정

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 유의확률을 이용한 가설검정
2. 양측검정과 단측검정
3. 마지막 코멘트

유의확률을 이용한 가설검정

유의확률을 이용한 가설검정

아까와 비슷하지만 살짝 다른 방식으로 가설검정을 할 수도 있다.

- 가설과 현실의 관계는 다음의 네 가지 가능성으로 압축된다.

의사결정 \ 실제	참인 H_0	거짓인 H_0
H_0 기각 못함	옳은 결정	제2종 오류(β)
H_0 기각	제1종 오류(α)	옳은 결정

- 특히 주목할 부분은 제1종 오류(Type I error)이다.
- 제1종류를 완전히 제거하는 것은 불가능하므로 α만큼은 용납해야 한다(Why?).
- 이렇게 "귀무가설이 옳은데 이를 기각할 확률"을 유의확률(p-value)이라고 한다.

$$\alpha = P(\underbrace{\text{reject } H_0}_{\text{의사결정 부분}} \mid \underbrace{H_0 \text{ is true}}_{\text{실제 부분}})$$

유의확률을 이용한 가설검정

유의확률을 이용한 가설검정의 핵심은 이것이다!

- 만일 (1) 귀무가설이 옳은데 이를 기각할 확률(p-value)이 (2) 충분히 작다면 자신있게 귀무가설을 기각할 수 있다.
- "용납"하기로 결정한 유의확률(p-value)의 수준을 유의수준(significance level)이라고 부르고 α로 표시한다.
- 관례적으로 10%, 5%, 1%, 0.1% 유의수준 중 하나를 많이 쓴다.
- 각각에 대응한 것이 90%, 95%, 99%, 99.9% 신뢰수준(confidence level)이다.

유의확률을 이용한 가설검정

유의확률이 충분히 작으면 자신있게 귀무가설을 기각한다.

- 예제 2를 다시 풀어보자($\mu_0 = 100, \bar{x} = 97, S = 10, n = 100$). 다만 유의수준을 5%로 설정하자(=95% 신뢰수준).
- 귀무가설이 옳다는 전제 아래 고래가 구한 표본평균을 표준화하면 다음과 같다:

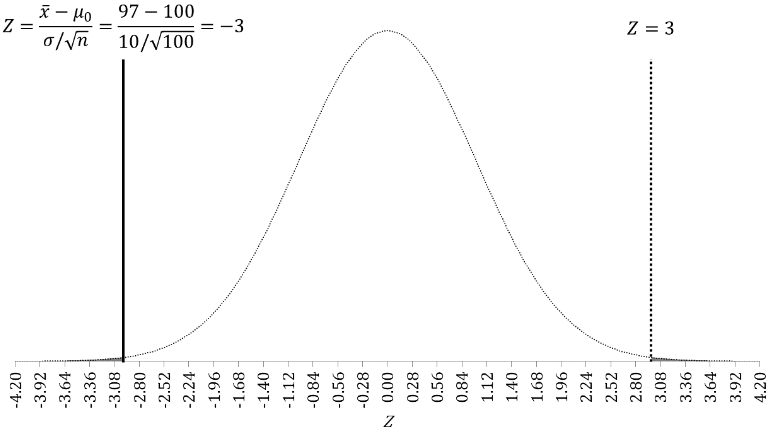
$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{97 - 100}{10/\sqrt{100}} = -3$$

- 만약 귀무가설이 옳다면 95% 신뢰수준인 [-1.96, 1.96] 안에 뽑힐 수 있는 표본평균의 95%가 들어있어야 한다.

유의확률을 이용한 가설검정

- 고래의 표본평균은 -3이다. 귀무가설이 옳다는 전제 아래에서 이 표본평균 내지 이보다 더 "극단적인" 표본평균이 나올 확률은 NORM.DIST(-3, 0, 1, TRUE), 즉 0.001 밖에 되지 않는다.
- 그런데 이게 전부가 아니다. 표본평균이 3 또는 이보다 더 "극단적인" 표본평균이 나올 확률인 1-NORM.DIST(3, 0, 1, TRUE)도 계산해서 "더해야" 한다.
- 왜냐하면 고래의 귀무가설은 양측검정 $H_0 : \mu = 100$이기 때문에, 귀무가설은 100보다 (1) 훨씬 작거나 (2) 훨씬 크다는 두 경우로 기각될 수 있기 때문이다.
- 그러므로 두 구간, NORM.DIST(-3, 0, 1, TRUE)과 1-NORM.DIST(3, 0, 1, TRUE)를 합한 값이 "귀무가설이 옳은데 기각할 오류 크기"인 α가 된다.

유의확률을 이용한 가설검정



유의확률을 이용한 가설검정

- 이 α 값이 5%보다 작다면 안심하고 귀무가설을 기각할 수 있다.
- 왜냐하면 이렇게 기각할 때 고래가 무릅써야하는 위험, 즉 "귀무가설이 옳은데 기각할 확률(p-value)"이 5%보다도 작기 때문이다.
- 예제 2에서 유의확률은 0.002이므로 (기준이 되는 유의수준인) 0.05% 보다 작다.
- 이렇게 유의확률이 유의수준보다 큰지 작은지 비교하여 즉각 가설을 검정할 수 있다.

유의확률을 이용한 가설검정

인정할 것은 인정하고 넘어가자.

- 근데 -1.96보다 작은 Z-값들과 1.96보다 큰 Z-값들로 채워지는 나머지 5%도 여전히 (낮을지언정) “귀무가설이 옳을 확률”을 보여주기는 한다(Why?).
- 다시 말해, 귀무가설이 옳더라도 5%의 확률로 [-1.96, 1.96] 바깥에서 표본평균이 뽑힐 수야 있다.
- α는 오류의 크기를 보여주며 이를 0으로 만드는 것은 영원히 불가능하다(Why?).

유의확률을 이용한 가설검정

예제 3. 청주에서 초등학교 1학년용 표준화 읽기검사 도구는 평균점수 100점, 표준편차는 15점을 갖도록 개발되어 있다. 브래드는 교육사회학적 관점에서 조기교육에 관한 연구보고서를 쓰고 있다. 그는 자녀교육에 열성인 부모를 가진 아이들이 입학 전 조기교육을 받기 마련이고, 그 때문에 읽기검사 도구의 당초 기대와는 다른 점수가 나오지 않을까 생각하고 있다. 브래드는 이 연구가설을 테스트해보기 위해 조기교육을 1년 이상 이수한 초등학교 1학년생 49명을 표본으로 뽑아 표준화 읽기검사를 실시하였다. 그 결과 표본의 평균점수는 105점이였다. 브래드의 적절한 가설을 제시하고 5% 유의수준에서 검정하시오.

유의확률을 이용한 가설검정

- 브래드의 가설은 다음과 같다(Why?).

$$H_0 : \mu = 100$$
$$H_a : \mu \neq 100$$

- (귀무가설이 옳다는 전제 아래에서) 평균이 100점인 모집단으로부터 무한히 계속 표본평균들을 구해 (가상적인) 표집분포는 정규분포할 것이다(Why?).
- 주어진 유의수준이 5%라는 것은 "귀무가설이 옳은데 이를 기각할 확률"로 5%까지 용납할 의향이 있다는 뜻이다.

유의확률을 이용한 가설검정

- 한편 (귀무가설이 옳다는 전제 아래) 브래드가 구한 표본평균을 표준화하면 다음과 같다.

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{105 - 100}{15/\sqrt{49}} = 2.333$$

- 귀무가설이 옳은데 표본평균이 -2.333보다 작은 값이 나올 확률은 NORM.DIST(-2.333,0,1,TRUE)로 구한다.
- 귀무가설이 옳은데 표본평균이 2.333보다 큰 값이 나올 확률은 1-NORM.DIST(2.333,0,1,TRUE)로 구한다.
- 두 면적은 각각 0.0098로 모두 더하면 0.0196다.

유의확률을 이용한 가설검정

- 5% 유의수준은 “귀무가설이 옳은데 기각할 확률”이 5%라는 의미이다. 이런 상황에서 귀무가설을 기각할 때의 위험은 제법 낮은 것이다(Why?).
- 그런데 브래드가 실제로 구한 유의확률은 1.96% 였으므로 (귀무가설을 기각할 때의 위험이) 5%보다도 낮은 셈이다.
- 그러므로 브래드는 5% 유의수준에서 (또는 95% 신뢰수준에서) “통계적으로 유의하게” 귀무가설을 기각할 수 있다.
- “조기교육을 받은 이들의 읽기점수는 그렇지 않은 이들의 읽기점수와 다르다”고 결론내릴 수 있다.

유의확률을 이용한 가설검정

임계값을 사용하는 방법과 유의확률을 사용하는 방법 둘 다 이해해야 한다.

- 만약 컴퓨터를 사용할 수 없고 오로지 수계산만 해야 한다면 임계값을 사용하는 방법이 편하다.
- 하지만 실무에서나 연구에서 압도적으로 유의확률(p-value)을 사용하는 방법이 선호된다.
- 컴퓨터 통계분석 패키지를 사용하면 자동적으로 p-value를 알려주는데, 여러분은 단지 그 값이 0.01 (10% 유의수준), 0.05 (5% 유의수준), 0.01 (1% 유의수준) 보다 작은지 쓱 훑어보면 유의성 여부를 바로 알 수 있기 때문이다.
- 시험에서는 어쩐지 유의확률을 사용하는 방법만 나올 것 같은 예감이 든다(아님 말고).

양측검정과 단측검정

양측검정과 단측검정

단측검정은 (아까 배운) 양측검정보다 오히려 더 쉽다.

- 다음과 같은 구조의 가설을 테스트했기 때문에 지금까지 배운 가설검정 방법은 사실 모두 양측검정(two-tailed test)이다.

$$H_0 : \mu = \mu_0 \qquad H_a : \mu \neq \mu_0$$

- 이와는 달리 다음과 같은 가설이라면 단측검정(one-tailed test)을 사용해야 한다.

$$H_0 : \mu \geq \mu_0 \qquad H_a : \mu < \mu_0$$

또는

$$H_0 : \mu \leq \mu_0 \qquad H_a : \mu > \mu_0$$

양측검정과 단측검정

- 예제 2를 살짝 바꾸어 여전히 $\bar{x} = 97$, $s = 10$, $n = 10$이지만 가설이 다음과 같다고 하자.

$$H_0 : \mu \geq 100$$
$$H_a : \mu < 100$$

- 이 때 표본평균을 표준화한 값은 변함없이 똑같다.

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{97 - 100}{10/\sqrt{100}} = -3$$

- 다만 색칠공부의 내용은 달라진다. 여기서는 $Z = -3$보다 왼쪽에 있는 끄트머리 면적은 곧 (모평균이 100 이상이라는) "귀무가설이 옳다는 전제 아래" 그런 극단적인 표본평균이 나올 확률을 보여준다.

양측검정과 단측검정

- 만일 이 면적이 (가령) 5%보다 작다면 "귀무가설이 옳은데 기각할 오류"가 충분히 작다는 의미이므로 곧바로 귀무가설을 기각할 수 있다.
- 물론 귀무가설이 옳다는 전제 아래 $Z = -3$보다 작은 값이 "우연히" 표본으로 추출되었을 수는 있다. 다만 그럴 가능성(=유의확률)이 매우 낮을 뿐이다.
- 양측검정의 경우에는 좌우 끄트머리의 유의확률(p-value)을 더해서 그것이 유의수준 α보다 작았을 때 귀무가설을 기각한다.
- 단측검정의 경우에는 한쪽 끄트머리의 유의확률(p-value)만을 보고 그것이 유의수준 α보다 작았을 때 귀무가설을 기각한다.

양측검정과 단측검정

- 만일 문제가 5% 유의수준으로 가설검정을 하도록 요구하였는데, 유의확률(p-value)이 심지어 0.01보다 작다면 어떨까?
- 물론 그냥 “5% 유의수준에서 귀무가설을 기각할 수 있다” 말해도 된다.
- 하지만 기왕이면 “1% 유의수준에서 귀무가설을 기각할 수 있다” 말하는 편이 더욱 좋다.
- 왜냐하면 이쪽이 더 엄격하고 보수적인 테스트조차도 통과했다는 의미를 전달하기 때문이다(“귀무가설이 옳은데 기각할 확률”이 더더욱 적었다!).

예제 4. 제천의 한 고등학교에서 고교생을 대상으로 교사에 대한 태도 검사 (200점 만점)를 실시한 결과 평균값은 178점이었고 표준편차는 34점으로 나왔다. 그 소식을 전해듣고 청주의 교육감 로즈는 청주에서 자신만만하게 36명의 임의표본을 추출하여 동일한 검사를 실시하였다. 청주에서 표본의 평균값은 170점이었고 표준편차는 24점이었다. 로즈는 풀이 죽어 이제 제천보다 정말로 청주의 고교생이 교사를 덜 존중하는가를 살펴보려고 한다. 적절한 가설을 제시하고 5% 유의수준에서 가설검정하시오.

양측검정과 단측검정

- 로즈의 가설은 다음과 같다(Why?).

$$H_0 : \mu \geq 178$$
$$H_a : \mu < 178$$

- 평균이 178점인 모집단에서 무한히 많은 표본평균들을 구해 (가상적인) 표집분포를 그린다면 이것은 정규분포할 것이다(Why?).
- 로즈의 표본평균을 표준화하면 다음과 같다.

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{170 - 178}{24/\sqrt{36}} = -2$$

양측검정과 단측검정

- $Z = -2$보다 왼쪽 끄트머리 음영의 면적이 가지고 있는 의미를 잘 고민해보면, 이것은 (귀무가설이 옳다는 전제 아래) Z값이 -2보다 작은 값이 나올 확률을 의미한다. 이것이 크다면 귀무가설을 기각할 수 없다(Why?).
- (귀무가설이 옳다는 전제 아래 그려진) 표집분포에서 표본평균이 -2보다 작은 값이 나올 확률은 NORM.DIST(-2,0,1,TRUE)로 구할 수 있다. 그 값은 0.023 정도로 5% 유의확률보다 작다.
- 계산된 0.023이라는 p-value는 실제 의미로 생각해 볼 때 “귀무가설이 옳은데 기각할 확률”이 겨우 약 2.3% 정도라는 것으로 100번 중 2.3번 정도만 오류를 저지른다는 것이다.
- 로즈는 5% 유의수준에서 (또는 95% 신뢰수준에서) 통계적으로 유의하게 “청주의 고교생은 제천의 고교생보다 교사를 같은 수준으로 또는 더 존경한다”라는 귀무가설을 기각할 수 있다.

양측검정과 단측검정

예제 5. 굿데이 타이어에서 생산중인 타이어의 평균 수명은 37,000km 이고 표준편차는 5,000km인 것으로 알려져 있다. 이 회사의 연구원 송이는 기존 공정을 혁신적으로 뒤바꾸는 쾌거를 달성하였다. 그녀의 비법에 따라 생산된 타이어 100개의 표본을 뽑아 조사한 결과 평균 수명은 무려 38,000km였고 표준편차도 동일하게 유지되는 것으로 알려졌다. 그녀의 혁신을 평가하기 위해 적절한 가설을 제시하고 5% 유의수준에서 이를 테스트하시오.

양측검정과 단측검정

- 송이의 가설은 다음과 같다(Why?).

$$H_0 : \mu \leq 37$$
$$H_a : \mu > 37$$

- 평균이 37인 모집단에서 무한히 많은 표본평균들을 구해 (가상적인) 표집분포를 그린다면, 그것은 정규분포할 것이다(Why?)
- 은비의 표본평균을 표준화하면 다음과 같다.

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{38 - 37}{5/\sqrt{100}} = 2$$

양측검정과 단측검정

- $Z = 2$보다 오른쪽 끄트머리 음영의 면적이 가지고 있는 의미를 잘 고민해보자!
- 이것은 (귀무가설이 옳다는 전제 아래) Z값이 2보다 큰 값이 나올 확률을 의미한다. 이것이 유의확률보다 크다면 귀무가설을 기각할 수 없다.
- 그런데 송이의 분석에서 대립가설(=연구가설)은 혁신제품의 평균 수명이 37,000km 보다 크다는 것이다(Why?).
- 그리고 오른쪽 끄트머리 음영이 바로 그러할 확률을 시사한다.
- 단측가설에서는 특히 부등식의 방향을 주의깊게 설정해야 한다! 최종적으로 귀무가설을 기각해놓고 귀무가설과 일치하는 결론에 도달해선 안된다.

양측검정과 단측검정

- 귀무가설이 옳은데 표본평균이 2보다 큰 값이 나올 확률, 즉 유의확률(p-value)은 1-NORM.DIST(2,0,1,TRUE)로 구할 수 있다. 그 값은 0.02275로 5% 유의수준보다 작은 값이다.
- 2.275%라는 유의확률은 실제 의미로 해석할 때 “귀무가설이 옳은데 기각할 확률”이 겨우 2.275% 정도라는 의미이므로 확신을 가지고 귀무가설을 기각할 수 있다.
- 송이는 5% 유의수준에서 (또는 95% 신뢰수준에서) 통계적으로 유의하게 새로운 공정이 기존의 공정과 같거나 그보다 못하다는 귀무가설을 기각할 수 있다.

양측검정과 단측검정

잘보면 귀무가설과 대립가설 중에 등호(=)는 항상 귀무가설 쪽에 붙는다.

- 양측가설이건 단측가설이건 H_0 쪽에 등호(=)가 붙어있다.
- 이는 귀무가설과 대립가설이 대립하는 최후의 순간(!)은 결국 등호(=)가 나온 부분이기 때문이다.
- 예컨대 아래처럼 "모평균이 50보다 크다"는 대립가설을 채택할 것인가를 결정한다고 하자.

$$H_0 : \mu \leq 50$$
$$H_a : \mu > 50$$

- 정말로 들여다보아야 하는 귀무가설은 "모평균이 50이 되는" 순간이고 "모평균이 50보다 작은" 부분들은 사실 별 의미가 없다(Why?).

양측검정과 단측검정

유의확률의 계산이라면 왼쪽 오른쪽 사이에서 너무 고민하지 않아도 된다.

- 표집분포는 $n > 30$의 조건 아래 결국 정규분포하고 이는 대칭적이다. 다시 말해, 양쪽 끄트머리의 면적은 같다.
- 하지만 문제를 성급히 풀려고 하기 전에 먼저 차분히 그림을 그리고 논리의 흐름을 되짚어 보자.
- 가설 구조에서는 올바르게 방향을 설정해야 한다.
- 계산하여 나온 유의확률(p-value)이 정말 “말이 되는지” 충분히 음미해 보아야 한다.

양측검정과 단측검정

우리는 오늘 하루종일 비율에 관한 가설검정을 다루지 않았다.

- 돌이켜보면 2주에 걸쳐 우리는 평균과 비율을 애써 구별하여 다루었다. 평균은 연속확률변수와 정규분포에, 비율은 이산확률변수와 이항분포에 대응시켰다.
- 그런데 오늘 평균에 관한 가설검정만 실컷 이야기했을 뿐이다. 그렇다면 비율에 대한 가설검정은 아무 의미도 없기 때문일까? 설마 그럴리는 없다.
- (표본 크기 n이 꽤 크다는 가정 아래) 비율 문제가 주어질 때 그냥 비율을 평균처럼 생각하고 풀면 된다.
- 어떻게 이것이 가능할까? 가장 근본적인 이유는 대규모 표본을 주로 분석하는 사회과학 분야에서 이항분포의 정규근사(normal approximation to the binomial)가 자연스럽게 활용될 수 있기 때문이다.

양측검정과 단측검정

- 앞서 나는 범주형 척도인 인종(1=백인; 2=흑인; ...)이나 종교(0=없음; 1=기독교; 2=불교; ...)의 평균에는 아무런 의미도 없다고 말했다.
- 하지만 여기에도 방법은 있다. 이런 범주형 척도도 일단 더미변수(dummy variable)로 만들면 비율의 의미가 생겨난다. 예컨대 "더미변수로서 백인"의 평균은 곧 백인의 비율을 반영하기 때문이다. 나중에 좀 더 자세하게 배우게 된다.
- 그렇기 때문에 (보건의료 통계학과는 별개로) 사회통계학에서는 비율에 관한 분석기법을 평균과 구분해서 배울 필요가 좀 적다.
- 물론 보건의료 통계학을 해야 하는 상황이라면 반드시, 그것도 꽤 잘 알아야 한다.

양측검정과 단측검정

교과서에 따라서는 이번 주 수업 내용은 제대로 다루지 않는다.

- 어떤 교과서에서는 이 파트에서 표준정규분포(=Z 분포) 대신에 t 분포(t distribution)와 단일표본 t 검정(one-sample t test)을 다루기도 한다.
- 왜 t 분포를 언급할까? 표본 크기가 작고 모표준편차를 모르는 경우에는 t 분포가 유일한 대안이기 때문이다.
- 우리는 바로 다음 주에 이 내용을 공부할 예정이다.

양측검정과 단측검정

- 특히 옛날 스타일로 쓰여진 교과서에서는 자꾸 표본 크기 n가 극단적으로 적은 상황과 결합하여 모표준편차 σ를 아는지 모르는지를 구분한다.
- 이런 구분 자체는 아예 현실성이 없다. 애시당초 모집단의 표준편차를 알 정도라면 표본을 구태여 뽑을 이유도 없다.
- 우리는 당연히 모표준편차 σ를 모르기 때문에 이것의 불편추정량(unbiased estimator)인 표본의 표준편차 s를 대신 사용하고 있다.
- 게다가 솔직히 말해 사례수가 30에도 미치지 않는 사회통계분석은 거의 없다.

마지막 코멘트

마지막 코멘트

가설검정은 기초사회통계를 마치고 중급사회통계로 진입하는 관문이다.

- 지금까지 걸어온 길을 반추해보자. 가설검정을 이해하기 위해서는 확률변수, 확률분포, 표집분포, 표준오차, 누적분포함수, 표준정규분포, 임계값, 유의수준을 모두 알아야만 했다.
- 이제 여러분은 가설검정을 이해하고 있다. 내가 본 대다수의 현업 사회조사 전문가는 이 이상의 통계 원리를 실무에서 사용하지 않는 것 같다(물론 미래에는 어떻게 될지 모른다).
- 이들의 노하우는 오히려 더 기본적인 것들을 잘 실천하는데 있다: 패널유지, 조사설계, 문항개발, 기본적인 표 만들기와 기술통계량 해석 등등!
- 믿지 못할지도 모르지만 여러분이 실무자로서 (어쩌면 연구자로서!) 가설검정을 사용해야 할 상황은 사실 굉장히 많다. 다만 스스로 그런 상황에 놓여있음을 눈치채지 못하는 것 뿐이다.

마지막 코멘트

그럼 사회통계의 남은 시간동안 우리는 무엇을 배울까?

- 가설검정을 할 수 있게 된 우리는 이제 그 원리를 활용하여 실천적인 기법을 연습한다.
- 두 개의 표본이 주어져 있을때 (1) 평균과 비율 그리고 (2) 분산을 비교한다. 몇 주 전에 살짝 맛보기 했던 (3) 교차표를 좀 더 공부한다. (4) 문항 신뢰도를 측정하기 위한 기법과 상관분석을 공부한다. (5) 회귀분석을 공부한다. 이렇게 남은 5주 동안 오로지 실천적인 기법만 연습한다.
- 가설검정은 물론 계속해서 쓰이게 될 것이다. 실천적인 기법 속에 가설검정이 어떻게 녹아들어가 있는지 알게 될 것이다.