사회통계

t 분포와 t 검정의 기초

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. t 분포와 자유도
2. 단일 모평균에 관한 가설검정

t 분포와 자유도

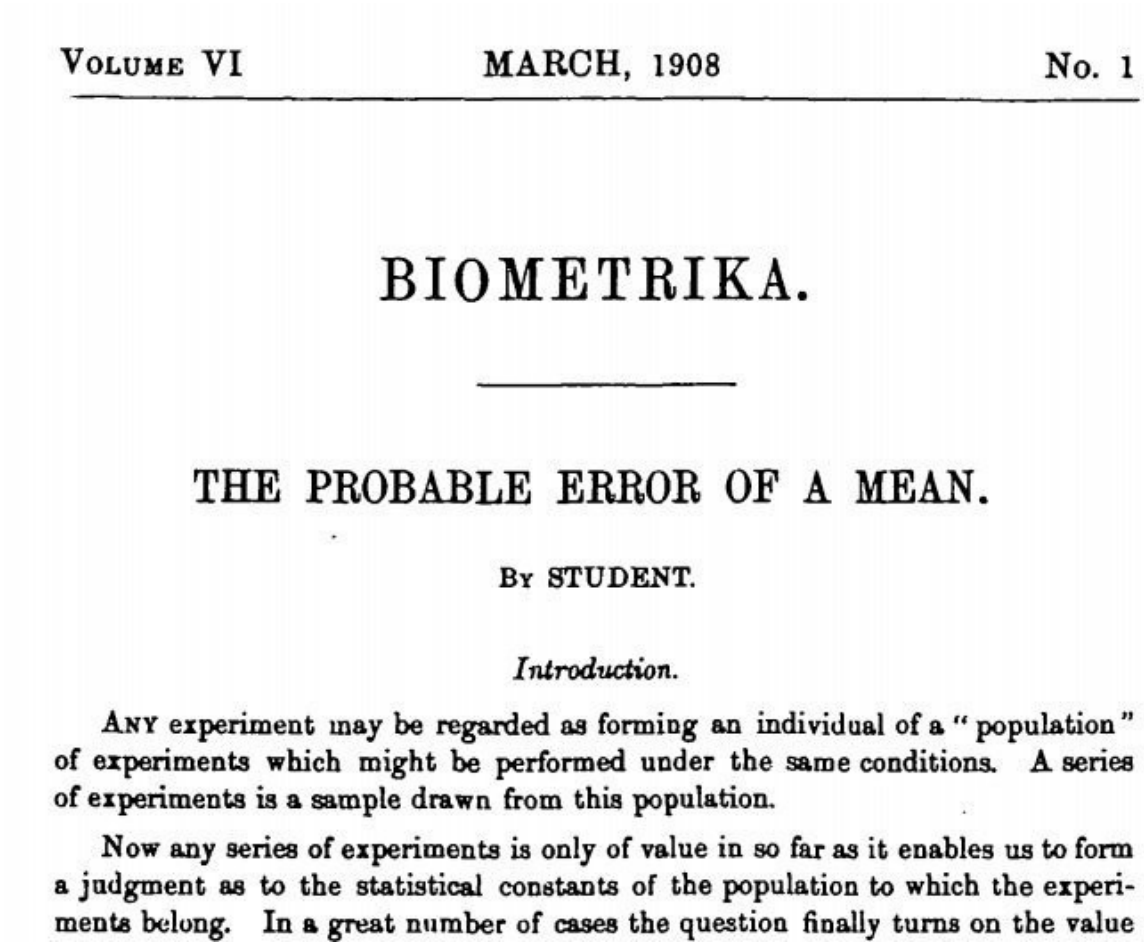
t 분포와 자유도

이론적 확률분포를 새로 하나 배우기로 하자.

- 몇 주 전에 우리는 이미 몇 가지 이론적 확률분포를 공부한 바 있다(e.g., 베르누이 분포, 이항분포, 정규분포 등).
- 오늘 배울 새로운 이론적 확률분포의 이름은 t 분포(t distribution)이다.
- 기본적으로 t 분포는 표준정규분포와 매우 비슷한 것임을 기억하자.
- 다만 (1) 사례가 적고(small Ns) (2) 모분산을 모르는(unknown variance) 경우라면 표집분포가 표준정규분포한다고 말하기 어렵다(Why?).
- t 분포는 이에 대응하여 고안되었다.

t 분포와 자유도

- t 분포는 1908년에 학생(Student)이라는 익명으로 출판된 논문에서 처음 알려졌다.



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value

t 분포와 자유도

새로운 분포를 사용하는데는 나름의 이유가 있다.

- 표본분산 s^2에 대해 잠깐 생각해보자(아래 식은 틀린 것이다).

$$s^2 = \frac{\sum (x-\mu)^2}{n}$$

- 원칙대로라면 모평균 μ을 사용해야 하지만, (모평균을 이미 아는 상황이 오히려 드물기 때문에) 우리는 표본평균 $\bar{x}$를 대신 사용할 수 밖에 없었다.
- 가만 생각해보면 이건 웃긴 것이다. 왜냐하면 지금 우리는 σ^2의 추정량(estimator)인 s^2를 구하려고 하는데, (진짜 들어가야 하는 μ 대신) 또다른 추정량인 $\bar{x}$를 사용했기 때문이다.
- 정말 이렇게 추정량을 이중으로 사용해도 문제는 없을까?

t 분포와 자유도

- 문제없을 리가 없다! 무엇보다 μ 대신 $\bar{x}$를 쓰면 표본분산 s^2는 (진정한 모분산인) σ^2 보다 작게 된다.
- 진정한 모분산 σ_x^2는 (1) 모평균 μ 대신 표본평균 $\bar{x}$를 사용해 추정된 표본분산 s_x^2와 (2) 표본평균 $\bar{x}$을 구하는 과정에서 발생한 표본평균의 분산 $\sigma_{\bar{x}}^2$의 합이기 때문이다 (Why?).

$$\sigma_x^2 = s_x^2 + \sigma_{\bar{x}}^2$$

- 즉 $\sigma_x^2 > s_x^2$ 이므로 (s_x^2로는 조금 부족한만큼) 이를 보정해줄 필요가 있는 것이다.

t 분포와 자유도

사실 William Gosset이 t 분포를 고안했다.

- Gosset은 (Ronald Fisher의 도움으로) 앞서 언급한 특수성을 인식하였고, 표본 수 n에서 "어쩔수 없이 사용한 $\bar{x}$ 같은 제한조건(limiting condition)"의 수를 빼면 이 문제를 해소할 수 있음을 발견했다.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- 분모 $n-1$ 부분을 자유도(degree of freedom; *df*)라고 부른다(교재 pp. 73, 169ff).
- 엑셀에서 STDEV.P(·) 함수와 STDEV.S(·) 함수를 사용했을 때 나타나는 미세한 차이는 바로 분모에서 표본 크기 n을 썼나, 아니면 자유도 $n-1$을 썼나의 차이에서 온 것이다.

t 분포와 자유도

이건 또 무슨 의미일까?

- 교과서에 따라서 자유도는 종종 "자료에서 자유롭게 바꿀 수 있는 관측치(observations)의 수"라고도 설명된다.
- 만일 $\bar{x} = 0$ 로 이미 정해져 있다면 $\left(\frac{-1+0+x}{3}\right)$ 의 상황에서 x에는 이미 자유가 없다(Why?). 따라서 $n = 3$이지만 $df = 3 - 1 = 2$이다.
- 잘 따져보면 표본 평균을 알고 있을 때, 실제로 새로운 정보를 제공하는 값은 n개 중 $n-1$개 뿐이다.
- 결론을 내려보자: "분모에서 표본 수 n 대신 자유도 $n-1$를 사용해야만, 표본분산 s^2는 모분산 σ^2의 불편추정량(unbiased estimator)이 된다."

t 분포와 자유도

t 분포를 사용하면 $n < 30$이라도 모평균에 관한 가설검정이 가능하다.

- 대표본을 가지고 모집단의 평균에 관한 가설검정을 한다면 자유도가 가져오는 변화는 몹시 작다(Why?).
- 대표본이라면 n이 크므로 거기에 1을 빼도 차이가 작다(예컨대 10,000에서 9,999로).
- 하지만 소표본(small sample)이라면 거기서 1을 뺐을 때 제법 체감이 크다(가령 10에서 9로).

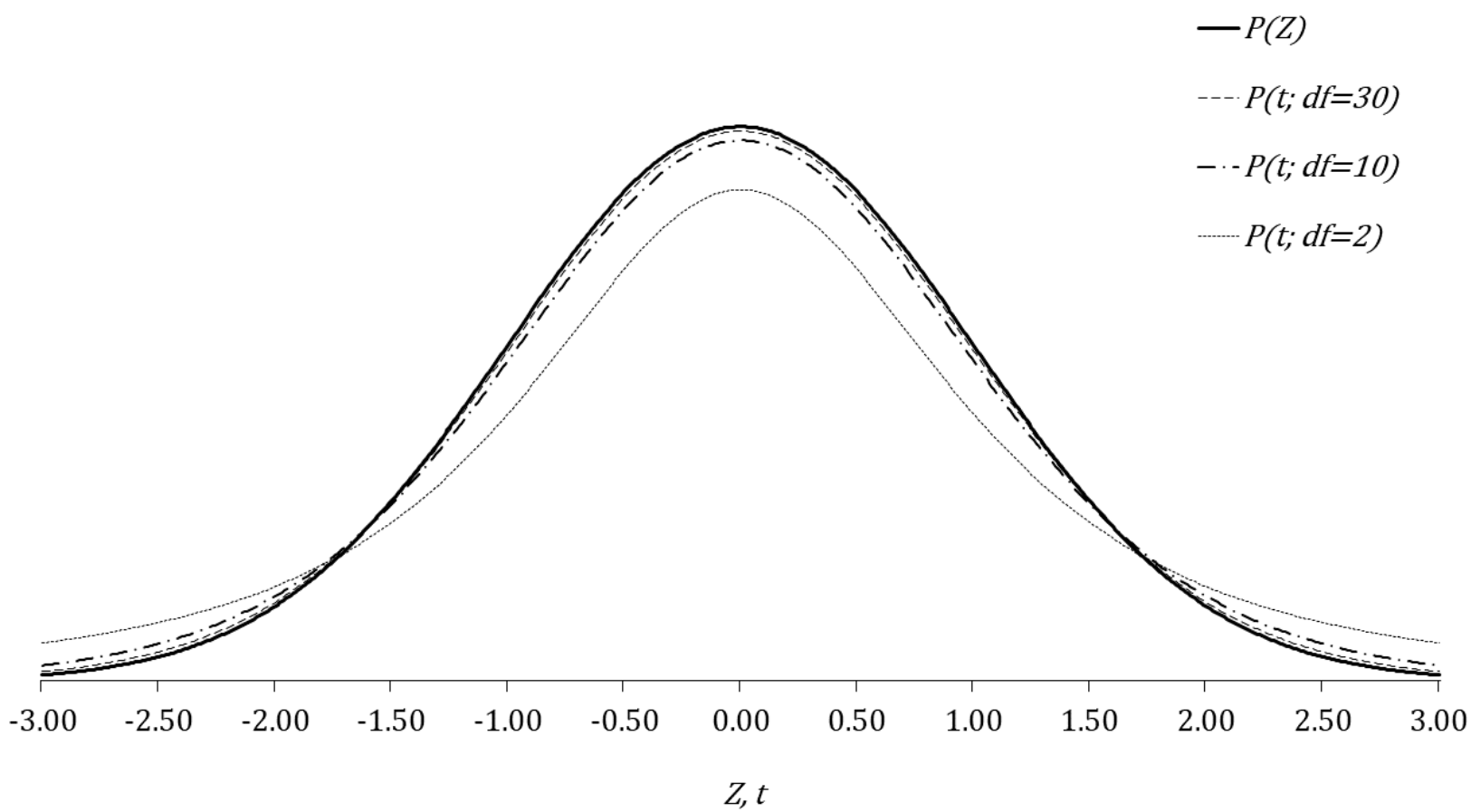
t 분포와 자유도

t 분포는 표준정규분포와 사실 매우 비슷하게 생겼다.

- (수학적으로 엄밀하게 다루지는 않겠지만) t 분포는 사실 Z 분포에서 파생된다.
- 정규분포의 패러미터(parameter)는 μ와 σ^2 였지만, t 분포의 패러미터는 오로지 자유도 df 뿐이다.
- 다시 말해, 표본의 μ나 σ^2에 의해 형태가 변하지 않고, 표준정규분포처럼 $\mu = 0$과 $\sigma^2 = 1$을 전제로 한다.
- 모평균을 추정하거나 가설검정하려는 경우에 제한조건은 $\bar{x}$ 딱 하나이므로 자유도는 $n - 1$이다.

t 분포와 자유도

- 자유도가 작을 때 t 분포는 표준정규분포에 비해 꼬리가 좀 더 두껍다!



t 분포와 자유도

- 대표본인 표집분포의 경우 다음의 공식에 따라 Z 값으로 변환할 수 있었다.

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- 소표본의 표집분포의 경우 다음의 공식에 따라 t 값으로 변환할 수 있다.

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

- 공식이 똑같으므로 표준화했을 때 Z 값이건 t 값이건 똑같다.
- 다만 어떤 분포를 사용하느냐에 따라 유의확률(p-value)이 달라질 뿐이다!

t 분포와 자유도

t 분포를 사용하여 (Z 분포와 비슷하게) 가설검정을 수행할 수 있다.

- 이제부터 공부할 t 검정(t test)에는 세 종류가 있다.
- 우선 살펴보려는 모집단이 하나 있을 때, 그 평균에 관해 가설검정할 수 있다.
- 모집단에 관한 귀무가설을 세우고 (이것이 옳다는 전제 아래) 그 표본평균의 표집분포를 (마치 중심극한정리처럼) t 분포로 표현할 수 있다. 그 상황에서 이러한 극단적인 값이 나올 확률을 계산하여 귀무가설을 기각할 수 있는지 판단한다.
- 이것이 (1) 단일표본 t 검정(one-sample t test)이다.

t 분포와 자유도

- 그 다음으로 살펴보려는 모집단이 두 개 있을 때, 그 평균의 차이에 관해 가설검정할 수 있다. 이른바 평균비교(mean comparison)이다.
- 두 모집단에서 각각 표본을 무한히 많이 뽑고 두 평균의 차이 $\bar{x}_1 - \bar{x}_2$를 확률변수로 판단하여 표집분포를 그릴 수 있다.
- 대표본이라면 중심극한정리에 따라 평균 차이의 표집분포는 정규분포한다.
  소표본이라면 평균 차의 표집분포는 t 분포한다.
- 이제 귀무가설을 세우고 적절하게 가설검정을 수행할 수 있다.
- 다만 자료구조에 따라 (2) 독립표본 t 검정(t test for independent samples) 또는 (3) 쌍체표본 t 검정(t test for paired samples)을 사용해야 한다.

단일 모평균에 관한 가설검정

단일 모평균에 관한 가설검정

단일표본 t-검정의 가설구조를 살펴보자.

- 다른 검정 기법의 가설구조와 마찬가지로 양측검정과 단측검정으로 나뉜다.
- 양측검정에서 귀무가설과 대립가설의 구조는 다음과 같다.

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

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

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

또는

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

단일 모평균에 관한 가설검정

예제 1. 생물정보분석팀의 고래는 급성골수성백혈병 환자들의 항암화학치료 이후 체중 변화를 연구하고 있다. 그러나 백혈병 환자는 워낙 드물어 임의표본으로 겨우 30명만을 추출할 수 있었다. 고래는 신중하게 문헌을 살펴본 뒤, "(chemotherapy를 마친 뒤) 평균 체중은 48kg일 것이다" 라는 귀무가설을 세웠다. 표본을 관찰해보니 막상 chemo를 마친 환자들의 평균은 44kg, 표준편차가 8kg였다. 고래의 귀무가설을 99% 신뢰수준에서 검정하시오.

단일 모평균에 관한 가설검정

- 사례가 매우 적으므로 (가상적인) 백혈병 환자 표본평균의 표집분포는 정규분포 t 분포를 따른다고 전제하는 것이 타당하다. 그 t 분포의 모양은 단순히 자유도에 의해 결정되므로 여기서는 30-1, 즉 $df = 29$다.
- 가설은 다음과 같이 설정할 수 있다.

$$H_0 : \mu = 48$$
$$H_a : \mu \neq 48$$

- Chemo 이후 환자들의 평균은 44kg, 표준편차가 8kg라고 했으므로 t 분포 위에 놓이는 t 값은 다음과 같다.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{44 - 48}{8/\sqrt{30}} = -2.7386$$

단일 모평균에 관한 가설검정

- 이제 고래는 (1) 귀무가설에 옳다는 전제 아래 그린 t 분포에서 (2) 주어진 표본의 평균같이 극단적인 t 값이 나올 확률(=유의확률)이 너무 적어서 (3) 귀무가설을 기각할 수 있는지를 판단해야 한다.
- 엑셀에서 NORM.DIST(·) 대신 T.DIST(·) 함수를 통해 t 분포의 "색칠 공부"를 할 수 있다. 이 함수에는 t 값, 자유도 df, (곡선 밑 면적을 구하므로) TRUE를 입력해야 한다.
- (문제에서 요구한대로) 1% 유의수준(=99% 신뢰수준)을 기준으로 하면, 엑셀에서 T.DIST(-2.7386, 29, TRUE) + (1-T.DIST(2.7386, 29, TRUE))를 계산하여 유의확률(p-value)을 구할 수 있다(Why?). 답은 약 0.0104이다.
- p-value가 0.01보다 살짝 크므로 고래는 귀무가설을 1% 유의수준에서 기각할 수 없다.

단일 모평균에 관한 가설검정

하지만 만일 Z 분포를 사용했다면 어땠을까?

- 사회통계 교과서를 다시 살펴본 고래는 자신의 샘플이 $n \geq 30$ 조건을 간신히 충족하므로, 중심극한정리에 따라 평균의 표집분포가 정규분포일지도 모른다고 생각하였다. 그러므로 이번에는 표준정규분포 상에서 가설검정을 다시 실시하였다.
- Z 값은 새로 계산할 필요조차 없이 t 값과 마찬가지로 -2.739이다(Why?).
- 엑셀에서 NORM.DIST(-2.739,0,1,TRUE)+(1-NORM.DIST(2.739,0,1,TRUE)) 함수를 활용하여 유의확률(p-value)을 구할 수 있다. 답은 약 0.006이다.
- p-value가 0.01보다 살짝 작으므로 고래는 귀무가설을 1% 유의수준에서 기각한다.

단일 모평균에 관한 가설검정

여기에서 t 분포와 표준정규분포의 결정적인 차이가 나타난다.

- 고래의 사례에서 확인할 수 있듯, 똑같은 표본을 가지고도 t 검정에서 기각하지 못했던 가설을 표준정규분포를 이용한 검정에서는 기각할 수 있는 상황이 생겨난다.
- 자유도가 작을때 t 분포는 표준정규분포에 비해 꼬리가 좀 더 두껍다.
- 이로 인해 t 분포를 사용하면 같은 신뢰수준(예컨대 90%, 95%, 99%)에 대해서도 좀 더 큰 검정통계량(test statistic), 즉 높은 t 값을 얻어야만 가설을 기각시킬 수 있게 된다.

단일 모평균에 관한 가설검정

- 왜 그런지 곰곰히 생각해보자!
- 소표본에서 얻어낸 검정통계량을 쉽사리 모평균이라고 신뢰할 수 있을까?
- 그럴수 없으니까 어지간히 큰 평균값이 나와주지 않는 한 쉽사리 귀무가설을 기각할 수 없는게 오히려 상식적이다.
- 이것을 보수적 추정(conservative inference)이라고 부른다.
- 물론 자유도가 커질수록(즉 표본 크기가 커질수록) t 분포는 표준정규분포와 점점 더 닮아가므로 보수적 추정의 수준이 점점 약해진다(Why?).

단일 모평균에 관한 가설검정

단일 모평균에 관한 가설검정과 관련하여 지금까지 두 가지 기법을 배웠다.

- 기본 원리와 엑셀 함수의 차이에도 좀 더 주목해야 한다.
- 대표본의 경우 NORM.DIST(x, mean, standard_dev, cumulative)를 사용했다. x에는 Z 값을 입력한다. 물론 표준화되었으므로 mean=0 그리고 standard_dev=1이다.
- 소표본의 경우 T.DIST(x, deg_freedom, cumulative)를 사용한다. x에는 t 값을 입력한다. (mean, standard_dev 대신) deg_freedom을 입력한다.
- 이때 결국 $t = Z$이다.
- "색칠 공부"이므로 둘 다 당연히 cumulative=TRUE이다.

단일 모평균에 관한 가설검정

예제 2. social.csv에서 socialself 변수는 초등학교 1학년생의 사회적 자아(social self)에 관한 여덟 개 문항에 대한 응답의 총합이다(코드북 참고). 새우는 "사회적 자아 점수의 평균은 32이다"라는 귀무가설을 세웠다. 적절한 가설 구조를 제시하고 5% 유의수준에서 이를 검정하시오.

단일 모평균에 관한 가설검정

- 모평균에 관한 가설검정이므로 적절한 가설은 다음과 같다.

$$H_0 : \mu = 32$$
$$H_a : \mu \neq 32$$

- 이 자료는 표본 크기가 매우 작다($n = 15$). 그러므로 표준정규분포가 아니라 t 분포를 사용해야 한다. t 분포는 자유도를 사용하며 이 자료의 자유도는 14이다.
- 엑셀의 [데이터] 메뉴에서 [데이터 분석]을 선택하고 "기술 통계법"을 고른다. 요약통계량을 구해 표본평균 $\bar{x}$과 표준오차 $s/\sqrt{n}$도 찾아낸다.
- 이제 t 값을 다음과 같이 계산한다.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{28 - 32}{5.41/\sqrt{15}} = -2.86$$

단일 모평균에 관한 가설검정

- 귀무가설이 옳다는 가정 아래에서 그린 표집분포 상에 이러한 표본평균 t 값이 나올 확률은 얼마인가? 다시 말해, 유의확률(p-value)을 확인해야 한다.
- 이 가설구조는 양측검정을 필요로 한다(Why?).
- 좌측 꼬트머리 면적은 T.DIST(-2.86, 14, TRUE)이고, 우측 꼬트머리 면적은 1-(T.DIST(2.86, 14, TRUE))이다.
- 유의확률은 약 0.0125이므로 5% 유의수준(=0.05)보다 작으므로 새우는 귀무가설을 5% 유의수준에서 기각할 수 있다.