가설검정

2023년 3월 16일 목요일 오전 11:45

통계적 추정

- 통계적 추정이란 부분으로 전체를 추리하는 것
  - 통계적 추정은 일상 생활에서 접하는 엄청나게 많은 데이터 세트 중에서 겨우 몇 개의 데이터를 관측하는 일에서 출발
  - 관측한 몇 개의 데이터로부터 그 뒷면에 펼쳐져 있는 엄청나게 많은 모든 데이터에 대해서 무엇을 관측할 수 있을까?
  - 관측된 데이터 뒷면에 펼쳐져 있는 모든 데이터를 통계학에서는 모집단이라고 부른다.
  - 통계적 추정은 관측된 데이터로 부터 모집단을 추리하는 것
- 정확한 모집단을 추정
  - 이상이 없는 N개의 동전으로 던지기 실험을 했을 때, 앞면이 10개 나왔다고 한다.
  - 던진 개수를 N으로 하여 다음가 같이 예상하는 것이 타당한지 않은지에 대해서 판단해보자
    1. 16개, 2. 36개
  - 추측하려는 N을 모집단이 가진 모수(Parameter)라고 부른다.
  - 여기서 모수는 예상하는 모집단의 종류에 대응하는 것이라고 이해하면 된다.

- 16개의 동전을 던져서 앞면이 나오는 개수를 예언한다면, 10개는 그 예언의 범위에 들어갈까?
  - 95% 예언 적중 구간를 계산해보자
  - 95% 예언적중구간은 부등식 $-1.96 \leq \frac{(x-\mu)}{\sigma} \leq +1.96$
  - 동전 던지기는 근사적으로 평균값이 $\frac{N}{2}$, 표준편차가 $\frac{\sqrt{N}}{2}$인 정규분포를 따른다.
  - 16개의 동전을 계산하면
    $\mu = \frac{16}{2}, = 8,$ 표준편차는 $\sigma = \frac{\sqrt{16}}{2} = 2$인 정규분포
  - 95%의 범위 안에 $-1.96 \leq \frac{x-8}{2} \leq +1.96$
  - $4.08 \leq x \leq 11.92$
  - 앞면이 나오는 개수는 4.08개 이상 11.92개 이하라고 예언
  - N=16이라는 상황에서 앞면의 10개 예측은 충분히 가능성이 있는 이야기
-
  - 36개의 동전을 계산하면
  - $12.12 \leq x \leq 23.88$
  - N이 36이라고 하면 현실에서 관측된 데이터 10은 예상할 수 없는 예상외의 수치
  - 이 때 가설 N=36을 타당하지 않다고 보고 버려버린다. 이를 통계학에서는 '가설을 기각한다.'라고 한다.
- 가설검정
  - 분산과 크기가 서로 다른 독립적인(unpaired) 표본들에 대한 t-검정
    - 모수적 검정의 일종
  - 맨-휘트니 U 검정
    - 비모수적 검정의 일종

가설

- 귀무가설(Null Hypothesis)
  - 통계에서의 가설 검정은 측정된 두 현상 간에 관련이 없다는 귀무가설(Null Hypothesis, H_0로 표시)
  - '관련이 없다'라는 형태의 가설
  - 두 변수가 독립이다. 두 변수의 평균에 차이가 없다. 동전을 던졌을 때 앞면이 나올 확률과 뒷면이 나올 확률에 차이가 없다. 특정 약이 질병 치료에 효과가 없다. 올해 제품의 생산량과 작년의 생산량이 같다.
  - 법정으로 비유하면 증거 불충분. 무죄추정의 원칙

- 대립가설(Alternative Hypothesis)
  - 두 현상간에 '관련이 있다'라고 보는 것으로 연구자가 알아보고자 하는 가설인 대립가설(Alternative Hypothesis; H_1로 표시)
  - 두 변수가 독립이 아니다. 두 변수의 평균에 차이가 있다. 동전의 앞면이 나올 확률이 동전의 뒷면이 나올 확률과 다르다. 특정 약]이 질병 치료에 효과가 있다. 올해 제품의 생산량과 작년의 생산량이 다르다.
  - 대립가설은 같지 않다. 작다. 크다 세 가지 형태로 나타낼 수 있다.
  - 대립가설은 귀무가설하고 다르게 피고가 유죄라고 판단

- 제 1종의 오류와 제 2종의 오류
  - 제 1종의 오류 : 귀무가설이 옳은데도 불구하고 그것을 버려 버리는 잘못
    - 실제로는 참이지만 연구결과 거짓이라고 나오는 경우
  - 제 2종의 오류 : 귀무가설이 잘못되었는데도 그것을 버리지 못한 잘못
    - 실제로는 거짓인데 연구결과 참으로 나오는 경우

가설검정

- 세포생물학을 공부하는 학생 100명을 50명씩 두 그룹으로 묶었다고 하자.
- 좀 더 큰 모집단에서 무작위로 선택한 학생들을 무작위로 두 그룹으로 나눈다. (그룹을 나눌 때 그 어떤 조건이 들어가지 않고 순수하게 무작위로 추출했다고 하자)
- 그룹 1은 강의를 듣고 실습과제를 수행
- 그룹 2은 강의만 들었다.
- 두 그룹의 기말고사 점수에 차이가 있는지 파악해 보자

표 4-1 그룹 1과 그룹 2의 시험 점수

그룹 1	81 80 85 87 83 87 87 90 79 83 88 75 87 92 78 80 83 91 82 88 89 92 97 82 79 82 82 85 89 91 83 85 77 81 90 87 82 84 86 79 84 85 90 84 90 85 85 78 94 100
그룹 2	92 82 78 74 86 69 83 67 85 82 81 91 79 82 82 88 80 63 85 86 77 94 85 75 77 89 86 71 82 82 80 88 72 91 90 92 95 87 71 83 94 90 78 60 76 88 91 83 85 73

- 두 그룹의 기말고사 점수에 유의미한 차이가 있는지 파악하려면 몇 가지 가설을 실험해 보아야 한다.
- 가설을 시험하는 데 사용하는 방법을 가설 검정(Hypothesis testing, 또는 가설 검증)이라고 한다.

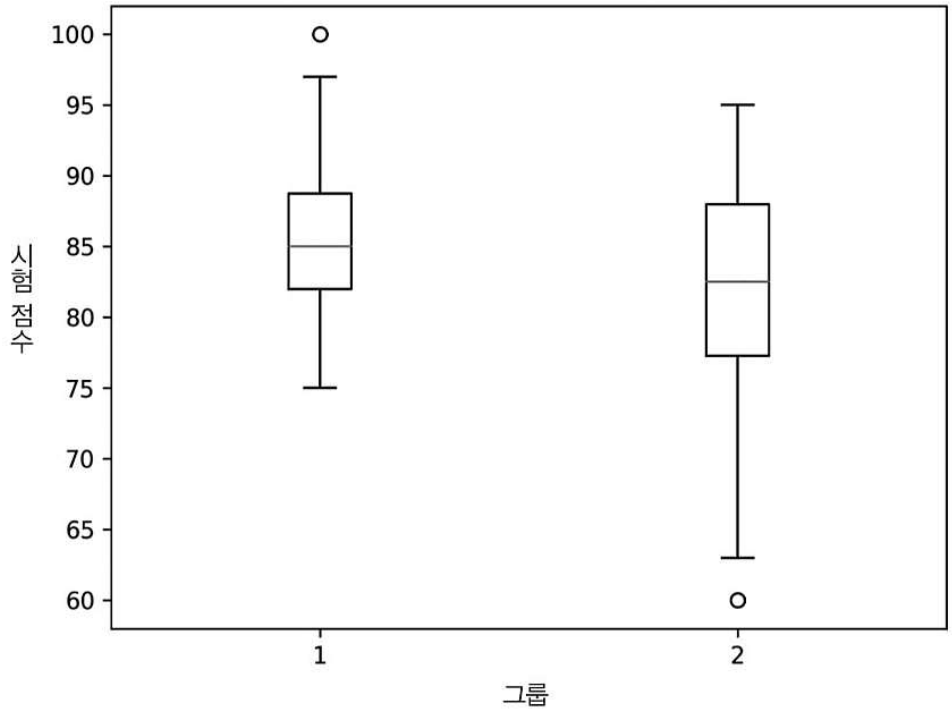


그림 4-7 표 4-1에 나온 데이터의 상자 그림

- 두 데이터의 집합의 평균들에 유의미한 차이가 존재하는가?
- 가설 검정으로 판단할 것은 두 데이터 집합이 같은 모집단에서 비롯했는지뿐이므로, 모든 검정은 양면(two-sided) 검정 또는 양측(two-tailed) 검정에 해당한다.

○ 접근 방식

1. 독립적인 두 데이터 집합을 비교하고자 한다.
2. 두 데이터 집합의 표준편차가 같은지에 대해서는 아무런 가정도 하지 않는다.
3. 귀무가설은 데이터 집합들의 모집단 평균들이 서로 같다는 것
   - 모집단 평균들을 알지 못하므로, 표본 평균들과 표본 표준편차들을 사용해서 승인할지 기각할지 판정하는데 필요한 증거를 얻는다.
4. 가설 검정은 데이터가 독립 동일 분포라고 가정한다. 위의 예제도 독립 동일 분포라고 가정

- t-검정
  - t로 표기하는 검정 통계량에 의존한다.
  - 미리 만들어진 t-분포와 비교해서 p-값을 구한다.
  - t-검정은 모수적 검정의 일종이다. 모수적 검정이라는 것은 검정할 데이터와 데이터의 분포에 관해 특정한 가정들을 둔다는 뜻

신뢰구간

- p-value와 함께 자주 등장하는 것이 신뢰구간(confidence interval CI)
- 신뢰구간은 우리가 비교햐는 두 데이터 집합들의 반복된 표본들의 평균 차이들이 일정한 비율로 속하게 되는 진 모평균 차이들의 구간
- 앞의 예제에서 앞면이 나온 개수가 10개로 관측될 때, 모수 N이 95%의 확률로 이 범위(13~30)에 들어간다라는 의미는 아님
- 동전 N개의 던져서 x개가 앞면이 된 경우, 이 x와 $\mu=N/2$, $\sigma=\sqrt{N}/2$ 로 부터 $z=(x-\mu)/\sigma$ 로 계산한 z가 부등식 $-1.96\leq z\leq +1.96$을 만족할 확률은 0.95이다.
- x를 관측하고 그 x에서 z를 계산해서 N을 기각해 가는 작업 한 경우, 정말 올바른 개수 N이 남을 확률은 각각의 관측값 x에 대해서 모두 0.95가 될 것이다. 그래서 어떤 관측값 x가 나온 경우에도 이 방법에서 N을 추정해 가는 과정을 반복한다면, 그 중 95%의 추정 결과는 맞다는 것이 올바른 해석이다.
- 다시 정리하면 95%라는 것은 구간 13~30에 정말 N으로 가능한 것이 95%로 들어간다라는 것이 아니라 '구간추정이라는 과정을 계속 실행하면, 관측값에 대응하는 여러 구간을 구할 수 있지만, 그 100번 중 95번은 N이 구해지는 구간에 들어간다'는 추정이 그 %가 된다.
- 95% 신뢰구간이란, 다양한 관측값에서 같은 방법으로 구간추정을 하면 그 중의 95%는 바른 모수를 포함하고 있는 구간을 말한다.

신뢰구간의 이해를 돕는 다른 표현

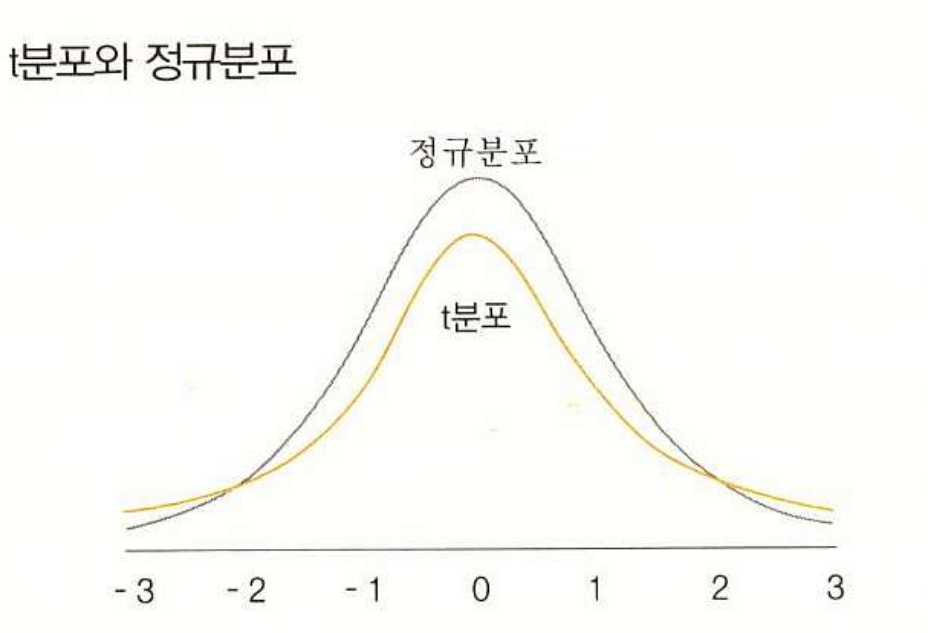
- 내가 95% 신뢰구간을 이용해서 100편의 논문을 사용했다면 평균적으로 5개의 논문에서 결론은 잘못되었다고 할 수 있고, 혹은 한 여론조사회사에서 100번의 여론조사(각가 다른 주제)를 했다면 이 중 (평균적으로) 5번은 신뢰구간이 참값을 포함하지 않는다고 설명할 수 있다.
- 불운하게도 예언 구간에 참 값이 포함되지 않을 확률이 5%나 되지만, 현재 내가 가지고 있는 구간이 그 100개 중 95개에 속한 것인지 5개에 속한 것인지 알 수 없다는 것

t분포를 이용한 정규모집단의 모평균 추정법

- 1단계
  - 얻은 n개의 표본에서 표본평균 $\bar{x}$와 표본표준편차 s를 계산한다.
- 2단계
  - 표본평균 $\bar{x}$와 표본표준편차 s, 추정하려고 하는 모평균 μ를 사용하여 자유도 n-1인 t분포를 따르는 통계량 T를 다음과 같이 계산
  - $T=\frac{(\bar{x}-\mu)\sqrt{n-1}}{s}$

- 3단계
  - 자유도 n-1인 95% 예언적중구간을 선택해서 $-\alpha \leq T \leq \alpha$라 하는 95% 예언적중구간을 만든다.
- 4단계
  - $-\alpha \leq \frac{(\bar{x}-\mu)\sqrt{n-1}}{s} \leq \alpha$를 μ에 대해서 풀면, 이것이 95% 신뢰구간이다.

t-분포와 정규분포



t-분포를 사용하여 모평균 추정 예제

어떤 가게 주인이 예상 매출액을 세우려고 한다. 주인은 매출액을 정규모집단에서 관측된 데이터로 가정하고, 이 모평균 μ를 대표적인 매출액으로 추정하려고 한다. 전표 중에서 무작위로 8장을 골라보니 다음과 같은 수가 나왔다.

- 45, 39, 42, 57, 28, 33, 40, 52 (만원)

모평균 μ의 구간을 추정해 보세요.

```
import scipy.stats as stats
import numpy as np

revenue = [45, 39, 42, 57, 28, 33, 40, 52]

# len(revenue) - 1의 자유도의 t-분포
target = t(len(revenue)-1)
print(f" 0.25 ~ 97.5 구간 -> {-target.ppf(0.975)} ~ {target.ppf(0.975)}")
print(f"표본평균 : {np.mean(revenue)}")
print(f"표본표준편차 {np.std(revenue)}")
```

0.25 ~ 97.5 구간 -> -2.3646242510102993 ~ 2.3646242510102993
표본평균 : 42.0

표본표준편차 8.831760866327848

- 휘발유의 옥탄가(정규분포로 가정함)를 13일 연속 조사하니 다음과 같았다.

  88.6, 86.4, 87.2, 88.4, 87.2, 87.6, 86.8, 86.1, 87.4, 87.3, 86.4, 86.6, 87.1

  옥탄가 모평균에 대한 95% 신뢰구간을 구하여라.

- 파이썬 코드

```
a = np.random.normal(85, 6, 50).astype('int32')
a[np.where(a > 100)] = 10
b = np.random.normal(82, 7, 50).astype('int32')
b[np.where( b > 100)] = 100
t, p = ttest_ind(a, b, equal_var=False)
print(("t = %0.5f, p = %0.5f") % (t,p))
```

t = 0.98472, p = 0.32722

- a는 평균이 85이며, 표준편차가 6.0인 정규분포에서 추출한 표본
- b는 평균이 82이며, 표준편차가 7.0인 정규분포에서 추출한 표본
- t는 검정 통계량, p는 계산된 p-value 값
- 0.05보다 작지만 0.05의 약 2분의 1 수준이지만, 귀무가설을 기각하고 두 그룹 a와 b가 서로 다른 모분포에서 비롯했다는 결론을 내리는 게 바람직함을 암시하는 증거이긴 함
- 실제로 두 표본은 서로 다른 정규분포에서 추출한 것이므로, 옳은 방향으로 점검 결과가 나옴 셈

- ttest_ind에서 ind는 independent 표본들에 사용하는 함수이다.
- equal_var = False도 두 표본의 분산이 같다고 가정하지 않는 웰치의 t 검정을 사용할 때는 이렇게 지정해야 한다. (위의 예제에서 두 표본의 분산이 같지 않음을 알고 있음)