6장 연습문제

문제 1. t-검정과 z-검정의 비교

5장에서 다룬 z-검정은 모집단의 모표준편차 값을 알고 있을 때 사용하는 반면, 이번 강의에서 배운 t-검정은 모표준편차 값을 모를 때 사용한다. 즉 t-검정은 z-검정에 비해 더 적은 정보만을 사용하므로, 신뢰구간이 더 넓을 것으로 유추할 수 있다. 실험을 통해 이를 확인해보도록 하자.

1-1) 아래 과정을 500회 반복하여, t-검정의 95% 신뢰구간이 z-검정의 95% 신뢰구간보다 큰 비율을 구하시오. Hint: 반복문 시행 전에 `count = 0`을 정의하고 t-검정의 신뢰구간의 너비가 더 넓을 경우 `count`값을 1씩 증가시키는 방식을 사용하라.

- 표본 수 n=30에 대하여, 표준정규분포 $N(0,1)$에서 n개의 표본을 추출하시오.
- 모표준편차 1을 알고 있다는 가정 하에 z-검정을 활용하여 95% 신뢰구간의 너비(신뢰상한-신뢰하한)를 구하시오.
- 모표준편차를 모른다는 가정 하에 t-검정을 활용하여 95% 신뢰구간의 너비를 구하시오.
- 두 값을 비교하고 t-검정에서 구한 값이 더 클 경우 `count`값에 1을 더하시오.

1-2) 1-1) 과정을 n값을 변화시켜 보면서 확인하시오. 구체적으로, n = 10,30,50,100,500,1000 에 대하여, t-검정의 신뢰구간이 z-검정의 신뢰구간보다 큰 비율을 구하시오. 이 결과가 갖는 의미는 무엇인가? Hint: $n \to \infty$ 일 때 $t(n-1) \to N(0,1)$ 이고 $S^2 \to \sigma^2$ 이다.

문제 2. 등분산성 검정의 이해

예제 2에서의 수업 전, 그리고 수업 후 시험성적 자료를 이용하여 아래 질문에 답하시오.

2-1) 두 집단의 모분산이 차이가 있는지 유의수준 5%에서 검정하시오.

2-2) 두 집단의 모분산이 차이가 있는지 유의수준 5%에서 검정하되, 2-1)와 집단의 순서를 반대로 하여 시행해보시오.

2-3) 2-1)과 2-2)의 결과를 비교하시오. 두 경우의 F 통계량은 어떤 관계가 있는가? Hint: F 분포의 성질을 참고하라.

문제 3. run10samp.txt

주어진 자료는 2012년 Washington, DC에서 열렸던 Cherry Blossom 10 mile run 경기에서 완주를 한 선수 100명의 자료이다. 주요 변수에 대한 설명은 다음과 같다.

변수명	설명
time	10 마일 달리기 완주 기록 (분)
age	선수 나이
gender	성별 (M=남성, F=여성)
state	출신 지역

3-1) run10samp.txt 자료를 읽어와서 run10samp 라는 이름으로 저장하고, 이를 사용하여 35살 이상과 35살 미만의 사람들의 완주 기록에 차이가 있는지 유의수준 5%에서 검정하시오.

In [ ]:

```
run10samp = pd.read_csv('run10samp.txt', sep=" ")
```

(아래부터는 별도의 등분산성 검정 없이 equal_var=False로 설정하시오.)

3-2) run10samp 에서 남성 참가자만 보았을 때, 35살 이상과 미만의 사람들의 완주 기록의 차이가 있는지 유의수준 5%에서 검정하시오.

3-3) run10samp 에서 출신 지역이 각각 'DC', 'MD', 'NY' 인 표본에 대하여, 각 지역별 표본의 갯수를 구하시오.

3-4) 3-3)에서 완주 기록의 평균이 가장 큰 지역은 'NY'이고 가장 작은 지역은 'MD'이다. 그 두 지역에 대해 지역별 완주 기록의 차이가 있는지 유의수준 5%에서 검정하시오. 유의확률 값은 얼마인가?

3-5) 3-4)에서 평균값이 가장 작은 'MD'지역과 가장 큰 'NY'지역을 비교하였으므로, 그 중간인 'DC'지역과 'MD'지역의 차이는 그보다 작을 것으로 예상할 수 있다. 한편, 유의확률은 귀무가설이 참일 경우 해당 사건이 발생할 확률이므로, 유의확률 값이 크면 귀무가설(두 집단의 모평균의 차이가 없다)을 채택할 근거가 강하다고 볼 수 있다. 'DC'와 'MD'에 대하여 지역별 완주 기록의 차이가 있는지 유의수준 5%에서 검정해보고, 유의확률 값을 3-4)와 비교하시오. 위의 예상대로라면 3-4)의 유의확률 값이 더 작은 것이 자연스럽다. 실제로는 어떠한가? 왜 이런 결과가 발생하는가? Hint: 3-3)의 결과를 참고하라