***0614** 이어서

+) 귀무가설 관련

-귀무가설은 모집단에 관련된 것이기 때문에 직접적으로 검정할 수 없음

*(*모집단은 직접 가지고 있는 데이터가 *X* <-> 표본은 가지고 있는 데이터*O)*

-대립가설은 표본에 관련된 것이기 때문에 직접적으로 검정이 가능함

-대립가설 기각 = 귀무가설 채택 / 대립가설 채택 = 귀무가설 기각

[엑셀과 파이썬을 이용한 통계 분석**]**

[6] 단일표본 **Z**검정

1. 단일표본 **Z**검정

<문제>

Q7. 실제 임계값이 나타내는 것은 무엇입니까?

Q.8 검정통계량(획득된 값)이 나타내는 것은 무엇입니까?

A. 검정통계량은 표본을 이용해서 계산한 값이다.

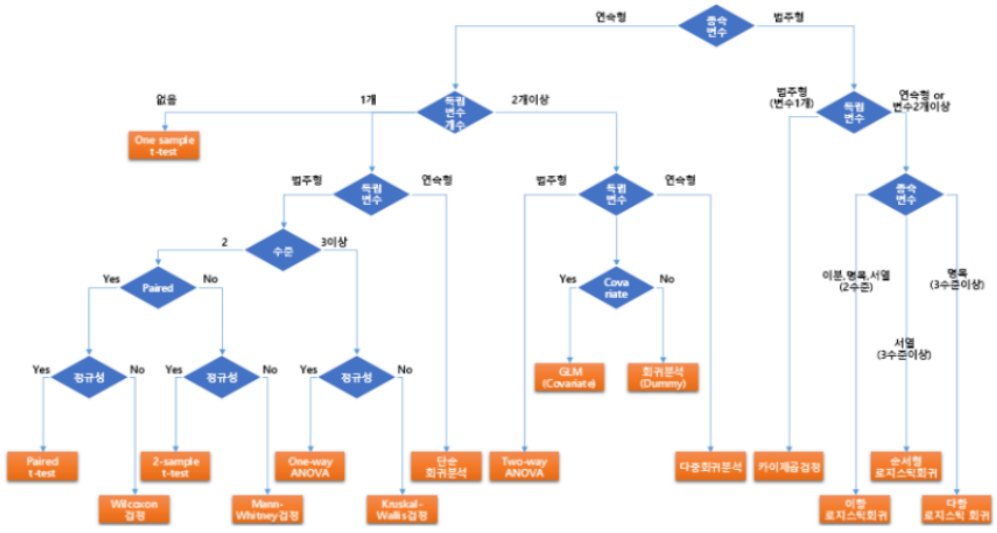
-유의확률 = 면적 = 확률

-검정통계량 = 유의확률에 해당하는 축의 값 = 우연히 발생할 확률에 해당하는 축의 값

-우리가 원하는 결과가 되려면 면적은 점점 작아져야 하고, 축의 값은 점점 커져야 한다.

Q9. 아래 순서도를 활용하여 2개의 독립된 집단 간의 차이점을 살펴보려면 어떤 단계를 가져야 합니까?

A.



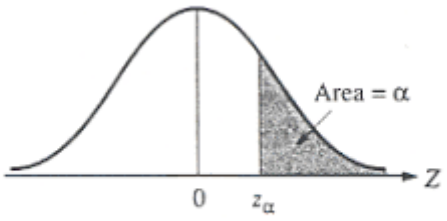
2) 관계 or 차이

3) 차이를 본다면, 짝지은 표본 or 독립표본

**우리가 선택해야 할 것은 '독립표본 t검정'

-두 개 이상의 표본 + 차이 + 독립표본&2개의 표본

Q10. 아래그림의 곡선의 오른쪽 부분에 색칠되어 있는 영역이 있습니다. 이 부분은 무엇을 나타냅니까?



<그림 9.2> 임계값 z_0

A. 유의수준 = 면적 = 확률

기각역 = 귀무가설 기각역

귀무가설 기각 = 대립가설 채택 = 차이가 있다는 내용

통계적으로 유의한 차이가 있다

유의확률이 낮다 = 우연히 발생할 확률이 낮다

Q11. 연구가설을 좀 더 엄격한 수준에서 검정한 경우(0.05에서 0.01로 이동), 그 영역은 더 커지나요, 아니면 더 작아지나요? 그 이유는 무엇입니까?

A. 유의수준이 0.05에서 0.01로 이동하면, 오른쪽으로 이동한다. 따라서, 유의수준이 작아지므로 면적은 작아진다.

---

2. 검정통계량 공식

$$Z_{stat} = \frac{\overline{X} - \mu}{SE}$$ (SEM) : 평균의 표준오차

-> 표준편차

2) 표본평균 **-** 모평균 **=** 표본오차 **->** 표준오차

$$SE = \frac{\sigma_X}{\sqrt{n}}$$: 표준오차(Standard Error)

표본표준편차 / σ : 모표준편차(σ 모르면 이 공식에서 S 사용하기)

* σ모르면 S 사용해야 하지만, σ를 사용하는 것이 좋음(=모수를 사용하는 것이 더 좋다)

*용어 정리

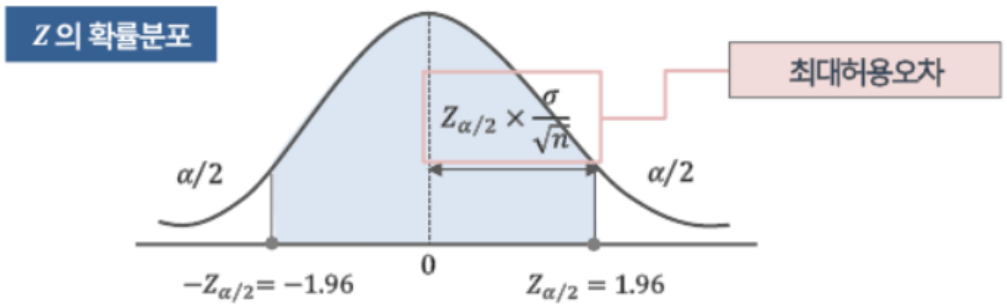
-SE = 표준오차 ≠ 표본오차(Sampling Error) = 오차한계

+) 표본오차 : 표본을 뽑을 때, 모집단을 온전히 대표할 수 있는 표본을 뽑는 데 실패할 오차

오차한계와 같은 개념임

(오차한계 : 추정 시, 모평균 추정구간의 중심으로부터 최대한 허용할 최대허용오차)

(즉, '표본오차를 구해라' 라는 문제를 보면, 오차한계를 구하면 됨 > 공식 : 임계값*SE)



3. 검정 과정**(**가설 검정**)** 순서

1) 귀무가설 및 연구가설의 진술

H0 : $\bar{X} = \mu$

H1 : $\bar{X} \neq \mu$

2) 귀무가설과 관련된 위험 수준(또는 유의수준 또는 1종 오류)

- 위험 수준 또는 제1종 오류 수준 또는 유의수준은 연구자가 결정함(ex. 0.05)

3) 적절한 검정통계의 사용(순서도 참고)

4) 검정통계량(획득된 값)의 계산

5) 특정 통계에 대한 적절한 임계값 표를 사용하여 귀무가설을 기각하는 데 필요한 값 결정

- ex. 1.96의 z값이 0.025의 확률과 관련되어 있음

6) 검정통계량과 임계값의 비교

- ex. 계산된 z값은 2.38이며 귀무 가설을 검정하기 위한 임계값은 +-1.96

7) 결론

*엑셀 활용하기

1) 수식 > 함수 더보기 > 통계 > Z.TEST

2) Array 선택(데이터가 포함된 셀 전부)

3) X 입력(이게 모집단에 속하는지 여부를 판별)

(+ 강사님 엑셀 파일 참고)

- 모분산을 알면 z분포 / 모르고 n>=30일 때도 z분포

- n<30이면 t분포

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ t통계량 공식

<문제>

Q1. 단일 표본 Z검정을 사용하는 것이 적절한 때는 언제입니까?

A. 단일표본이 주어졌을 때, 모집단과 표본의 차이를 알고 싶을 때

Q2. Z검정에서 Z는 무엇입니까? 단순 z점수 또는 표준 점수와의 유사점은 무엇인가요?

A. 검정통계량 z값

z점수와 유사하게 표준화시켜서 다른 분포와 비교가 가능하다

Q3. 다음의 상황에 대해 연구가설을 작성하세요.

Q3-1. 종현이는 초콜릿만 먹는 식이요법을 하는 그가 속한 집단의 체중 감소가 초콜릿만 먹는 식이요법을 하는 중년 남성 전체의 모집단을 대표하는지를 알고 싶습니다.

A. 표본평균(종현이가 속한 집단) ≠ 모평균(중년 남성 전체)

Q3-2. 보건복지부는 지난 독감 유행 기간 중 시민 1000명당 독감 발병 비율이 지난 50년 동안의 비율과 유사한지 비교해야 합니다.

A. 표본비율(지난 독감 유행 기간) ≠ 모비율(지난 50년)

Q3-3. 영규는 여러 아파트 건물의 주인입니다. 그는 현재 아파트에 대해 작년에 받은 월세가 지난 20년간의 월세를 대표하지 못한다고 거의 확신합니다.

A. 표본평균(작년) ≠ 모평균(지난 20년)

Q4. 서울 성동구에서 지난번 독감 유행 기간(4개월 또는 20주) 동안 발생한 독감 환자의 보고 건수는 일주일에 약 15건이었습니다. 서울 전체의 평균 보고 건수는 16건이었고 표준편차는 2.35였습니다. 성동구에 사는 아이들은 서울 전역의 아이들만큼 아플까요? 직접 계산해 보세요.

A. 표본 : 서울 성동구, 15개 / 모집단 : 서울 전체

-모표준편차와 모평균을 알고 있음 + n(15) < 30이므로 단일표본 z검정 실시

-z값 = (15-16) / (2.35 / √20) = -1.64808

-임계값 = 1.96

-유의수준 : 0.05

-양측검정(꼬리가 2개)

-유의확률 : 0.09

-결론 : 유의수준 > 유의확률 / 따라서 귀무가설을 채택하고 대립가설을 기각한다.

즉, 통계적으로 유의하지 않다.

++ 파이썬 코드로 작성하기

```
# 1) 모표준편차를 알고 있는 경우
n = 15;N = 16;x_bar = 15;mu = 16;sigma = 2.35

print(n,N,x_bar,mu,sigma)

# 검정통계량
# 분자  표본평균-모평균
x_bar - mu

# 분모  표준오차 = 모표준편차/제곱근(표본의 크기)
sigma/np.sqrt(n)

# 분자 / 분모
zv = (x_bar - mu) / (sigma/np.sqrt(n))
zv = abs(zv);zv
```

임계값

유의수준, 양측검정

alpha = 0.05

임계값에 해당하는 누적면적 = 1 - alpha/2 = 0.975

cv = norm.ppf(1 - alpha/2);cv

결론

cv < zv

왼쪽 = 채택역 = 귀무가설 채택

= 대립가설 기각 = 차이가 없다

= 통계적으로 유의한 차이가 없다

성동구에 사는 아이들은 서울시 전체 아이들만큼 아프지 않다.

유의확률

= 검정통계량에 해당하는 면적

pv = 1 - norm.cdf(zv);pv = pv * 2;pv

결론

alpha > pv

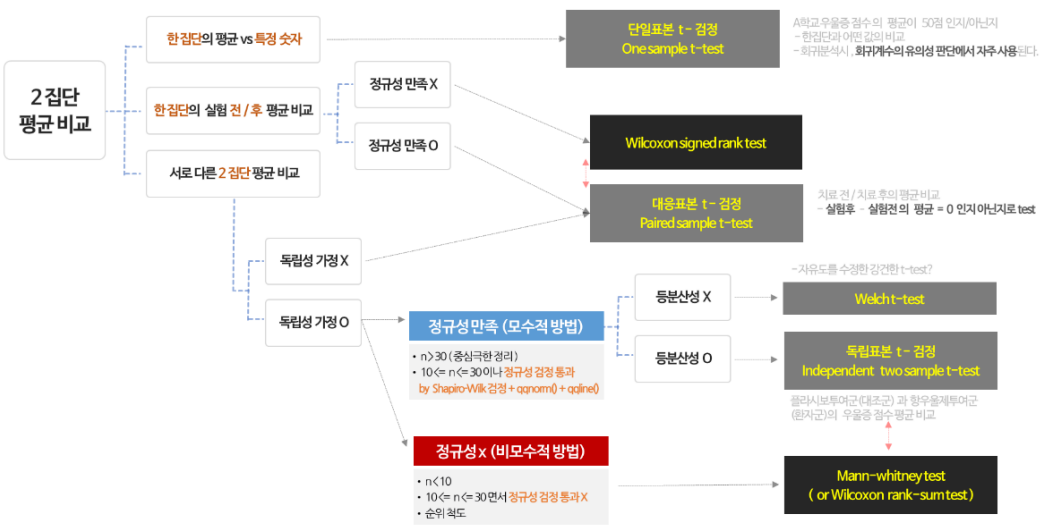
왼쪽 = 채택역 = 귀무가설 채택

= 대립가설 기각 = 차이가 없다

= 통계적으로 유의한 차이가 없다

[7] 독립표본 **t** 검정

1. 독립 표본 **t-**검정



(참고 : https://nittaku.tistory.com/467)

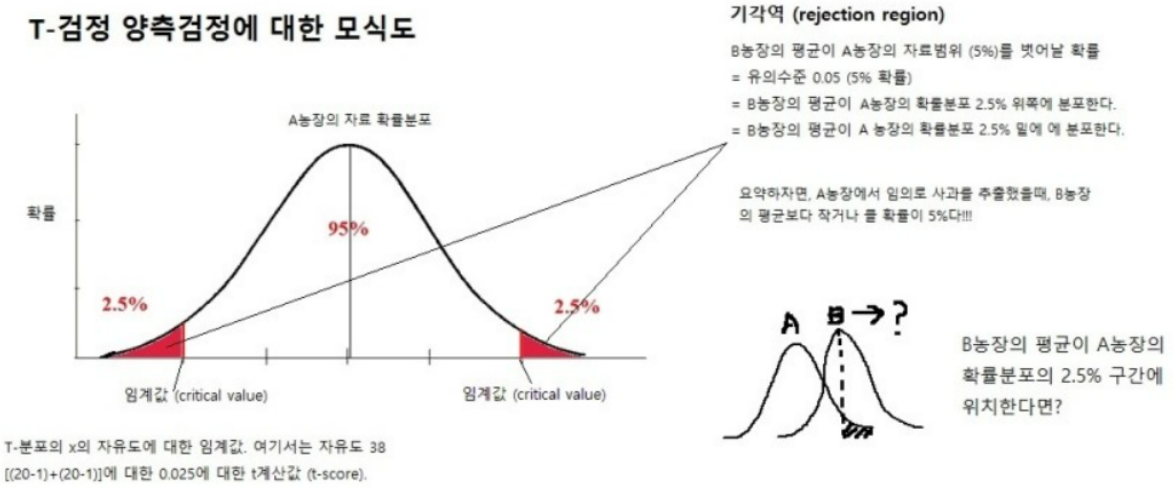
1) 검정통계량

-t값 계산에는 두 가지 방법이 있음 : 등분산 가정 / 이분산 가정

-t분포는 자유도가 필요함 (z분포는 누적확률로만 찾을 수 있음)

: 누적확률값과 자유도(n-1)가 만나는 지점을 구해야 함

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left[\frac{(N_1-1)S_1 + (N_2-1)S_2}{N_1+N_2-2}\right] \times \left[\frac{N_1+N_2}{N_1 N_2}\right]}}$$



두 사과 농장에서 계산된 t-값이 자유도 38로 만들어진 이론상의 t-분포표 상에 얼만큼의 확률로 존재하는가? 얼마만큼의 확률로 같을 수 있는가? 따라서 내가 정해놓은 확률보다 낮은가 높은가?
A농장과 B농장 간의 차이가 없는 확률을 얼마나 정할 것인가? (유의수준)

H0= A농장과 B농장간의 차이가 없다.
H1= 두 농장간의 차이가 있다.