

[엑셀과 파이썬을 이용한 통계 분석]

#### [4] 가설검정

##### 1. 가설

-분석(연구) 목적에 맞는 질문

-가설 검정은 표본을 이용한 것

##### 2. 표본과 모집단

-모집합 : 큰 집합, 전체 집합

-표본 : 작은 집합, 부분 집합

-분포에 대한 변동성 측정치

: 관측치 - 평균 = 편차 -> 표준편차

-표본과 모집단의 차이에서 발생하는 변동성 측정치

: 표본통계량 - 모수 = 표본오차 -> 표준오차

-표본이 모집단을 정확하게 대표할 때 연구 결과를 일반화할 수 있음

##### 1) 귀무가설(영가설 / null=zero의 의미)

: 연구(분석)의 출발점, 기준점

-차이가 없다 = 0

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 - \mu_2 = 0$$

-변수들 사이에 아무런 관계가 없다고 가정

Ex. 반응 시간과 문제 해결 능력 사이에는 아무런 **관련이 없다**.

##### 2) 대립가설(연구 가설)

: 연구(분석) 목적의 의미를 가짐

-차이가 있다  $\neq 0$

$$H1 : \bar{X}1 \neq \bar{X}2$$

$$H1 : \bar{X}1 - \bar{X}2 = 0$$

(H1 또는 Ha)

-변수들 간에 어떠한 관계가 있다는 확실한 진술

Ex. 반응 시간과 문제 해결 능력 사이에는 **관련이 있다.**

### 3. 양측검정

Ex. ABC 기억 검사에서 9학년의 평균 점수와 12학년의 평균 점수에는 **차이가 있다.**

$$H1 : \bar{X}9 \neq \bar{X}12$$

a. H1 : 첫 번째 연구가설

b.  $\bar{X}9$  : 9학년 표본의 평균 기억 검사 점수 /  $\bar{X}12$  : 12학년 표본의 평균 기억 검사 점수

c.  $\neq$  : 두 평균 기억 검사 점수가 같지 않다

### 4. 단측검정

Ex. ABC 기억 검사에서 12학년의 평균 점수가 9학년의 평균 점수보다 더 높을 것이다.

$$H1 : \bar{X}12 > \bar{X}9$$

### 5. 귀무가설과 연구 가설의 차이점

귀무가설	연구가설
차이가 없다 $= 0$	차이가 있다 $\neq 0$
모집단에 대해 언급 ( $\mu$ )	표본에 대해 언급 ( $\bar{X}$ )
간접적으로 검증	직접적으로 검증 실제로 분석할 수 있는 것은 표본 결과에 따라 연구가설을 채택 or 기각

## 6. 확률

-추론 통계의 기초

-정규분포에 사용

-유의수준 : 허용하는 오류의 정도 / 신뢰수준 : 확신의 정도

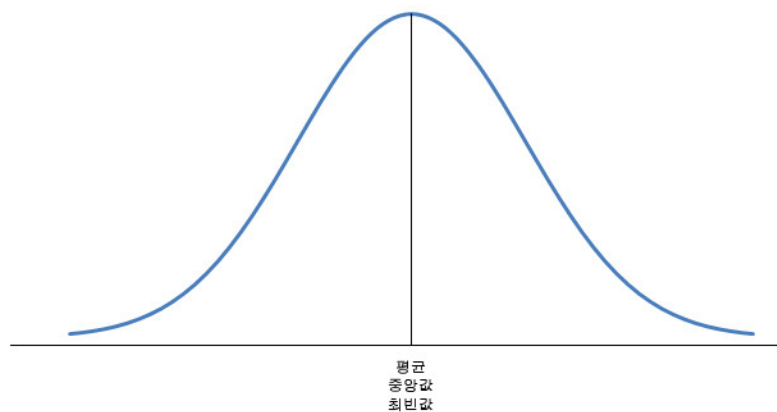
-전체 확률 = 1 = 신뢰수준 + 유의수준

Ex. 유의수준 0.05이면, 신뢰수준 =  $1 - 0.05 = 0.95$

## 7. 정규곡선(종 모양 곡선)

: 평균 = 중앙값 = 최빈값

-좌우대칭, 점근적 꼬리



## 8. 중심 극한 정리(CTL)

: 동일한 확률분포를 가진 독립 확률 변수  $n$ 개의 평균의 분포는  $n$ 이 적당히 크다면 정규분포에 가까워진다.

- $n$ 이 적당히 크다면의 기준 :  $n \geq 30$

\*\*여기부터 0614

[엑셀과 파이썬을 이용한 통계 분석]

[4] 가설검정

9. Z점수

: 표준점수(표준편차 단위로 표준화된 점수)이므로 서로 다른 분포 비교 가능

$$z = \frac{x - \mu}{\sigma}$$

-Z : 표준점수 / X : 개별 점수 /  $\bar{X}$  : 분포의 평균 / S : 분포의 표준편차

통계 용어	모집단	표본
평균	$\mu$	$\bar{X}$
표준편차	$\sigma$	S

<엑셀 활용>

-z점수 함수 : STANDARDIZE(X, 평균, 표준편차)

-누적확률(누적분포함수) 함수 : NORM.S.DIST(Z점수, True)

<파이썬 활용>

1) 수작업

```
[1] import pandas as pd

[2] x = pd.Series([12,15,11,13,8,14,12,13,12,10])

[3] # 평균
x_bar = x.mean();x_bar

12.0

[4] # 표준편차
s = x.std();s

2.0
```

```
[6] # z점수 = (관측치 - 평균)/표준편차
z_score = (x - x_bar)/s/z_score

0    0.0
1    1.5
2   -0.5
3    0.5
4   -2.0
5    1.0
6    0.0
7    0.5
8    0.0
9   -1.0
dtype: float64
```

2) 함수 이용 : StandardScaler()

```
[7] from scipy.stats import norm

[8] # 누적확률 = 누적분포함수
# 왼쪽에서 오른쪽으로 가면서 누적
norm.cdf(z_score)

array([0.5       , 0.9331928 , 0.30853754, 0.69146246, 0.02275013,
        0.84134475, 0.5       , 0.69146246, 0.5       , 0.15865525])

[9] # 오른쪽에서 왼쪽으로 가면서 누적
norm.cdf(-z_score)

array([0.5       , 0.0668072 , 0.69146246, 0.30853754, 0.97724987,
        0.15865525, 0.5       , 0.30853754, 0.5       , 0.84134475])
```

```
# 함수 이용해서 z점수 계산
# 1) 불면추정치 표준편차 이용
from scipy.stats import zscore
zscore(x, ddof = 1)

[11] # 아래 모델에 적용하기 위해 2차원 구조로 변경
x_df = pd.DataFrame({'x': x})

[12] # 2) 편향추정치 표준편차 값을 이용해서 z점수 계산
from sklearn import preprocessing
standard = preprocessing.StandardScaler()
standard.fit(x_df)
z_score = standard.transform(x_df)
z_score

zscore(x, ddof = 0)
```

## 10. 가설 검정과 z점수

-모든 사건에 그와 관련된 확률값이 존재

-사건이 발생할 확률(유의확률)이 얼마나 낮은지 확률값( $\alpha$ )을 이용하여 결정

-유의확률 : 우연히 발생할 확률 > 적으면 좋은 값( $\alpha = 0.05$ 보다 작으면 좋음)

Ex.

a. 유의수준( $\alpha$ ) > 유의확률(p) ::우리가 원하는 것!

-유의확률(우연히 발생할 확률)이 낮다 = 유의하다 = 차이가 있다

b. 유의수준 < 유의확률

-유의확률(우연히 발생할 확률)이 높다 = 유의하지 않다 = 차이가 없다

-근거가 약하기 때문에, 우리가 원하지 않는 결과

-통계적으로 유의한 자료가 없다

-z점수(z score) = 검정통계량 = 유의수준에 해당하는 **측의 값**

**\*\*여기서 1) 측의 값 2) 면적=확률 의 차이를 이해해야 함! >> 매우 중요**

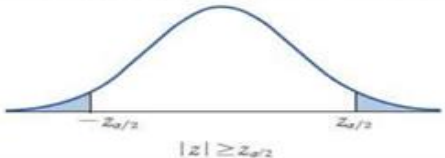

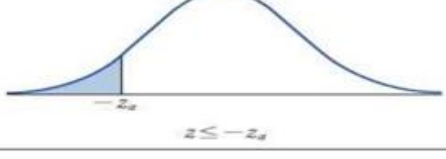
-1) 측의 값 : z점수

-2) 면적=확률 : 유의확률, 유의수준

-가설검정의 방법 2가지

① 유의확률과 유의수준 비교(**확률**)    ② 검정통계량과 임계값 비교(**측의 값**)

-양측검정과 단측검정 **\*\*우리가 원하는 건 오른쪽에 있다**

검정의 종류	귀무가설과 대립가설	기각역 (색칠한 부분의 가로축 좌표)
양측검정	$H_0: \mu = \mu_0$ 대 $H_1: \mu \neq \mu_0$	 $ z  \geq z_{\alpha/2}$
단측검정	$H_0: \mu = \mu_0$ 대 $H_1: \mu > \mu_0$	 $z \geq z_{\alpha}$
	$H_0: \mu = \mu_0$ 대 $H_1: \mu < \mu_0$	 $z \leq -z_{\alpha}$

(헷갈리는 개념 정리 참고 : <https://todayisbetterthanyesterday.tistory.com/4> )

## 11. 유의성의 정의

### 1) 통계적 유의성

: 우연히 발생할 확률이 낮다( $p < \alpha$ )

### 2) 세상에 완벽한 것은 없다

-유의수준( $\alpha$ ) : 오류를 수용할 수 있는 우연 또는 위험 수준

+ 신뢰수준 : 보통 95% (5%의 유의수준은 95%의 신뢰수준임)

\*유의수준, 신뢰수준 모두 면적임(확률)

## 12. 오류

### 1) 명백한 오류인 경우

-귀무가설이 실제로 참인데도 귀무가설을 기각

-귀무가설이 실제로 거짓인데도 귀무가설을 채택

### 2) 제1종 오류

=  $\alpha$  = 유의수준

-귀무가설을 검정할 때 연구자가 감수하고자 하는 오류나 위험의 수준

-일반적으로 0.05\*, 0.01\*\*, 0.001\*\*\*로 설정 (일반적으로 \*의 수로 위험도를 표시)

Ex.  $\alpha=0.05$  : 연구자가 귀무가설이 참일 때 귀무가설을 기각할 확률이 5%라는 의미

:  $p < 0.05$ 로 표시 / 이러한 결과를 관찰할 확률이 0.05 미만

통계적 결정 \ 실제상황	$H_0$ 가 사실 (참)	$H_0$ 가 허위 (거짓)
	$H_0$ 가 사실 (참)	$H_0$ 가 허위 (거짓)
$H_0$ 채택	옳은 결정 확률 = $1 - \alpha$	제II종오류 확률 = $\beta$
$H_0$ 기각	제I종오류 확률 = $\alpha$	옳은 결정 확률 = $1 - \beta$

검정력

유의수준

영가설(H0): 개발한 “신약은 효과가 없다”		실제상황	
		신약 효과 없으면	신약 효과 있으면
의사결정	채택 (계속 연구)	정상 업무	감봉 ( $\beta$ )
	기각 (신약 출시)	해고! ( $\alpha$ )	표창과 포상( $1-\beta$ )

(기본적으로 치명적인 경우 :  $\alpha$ (1종 오류) >  $\beta$ )

3) 통계적으로 유의하다 = 연구가 매우 성공적

-귀무가설 기각 = 연구가설 채택

-통계적으로 유의한 것이 항상 좋은 것은 아님

-통계적 유의성이 개념적으로 중요한 것은 맞지만, 최종 목적/목표는 아님!

-주관적인 상황에 따라 판단하기

## [5] 추론 통계

### 1. 추론의 원리

-추론 : 표본으로 모집단을 추론하는 것

1) 표본 추출

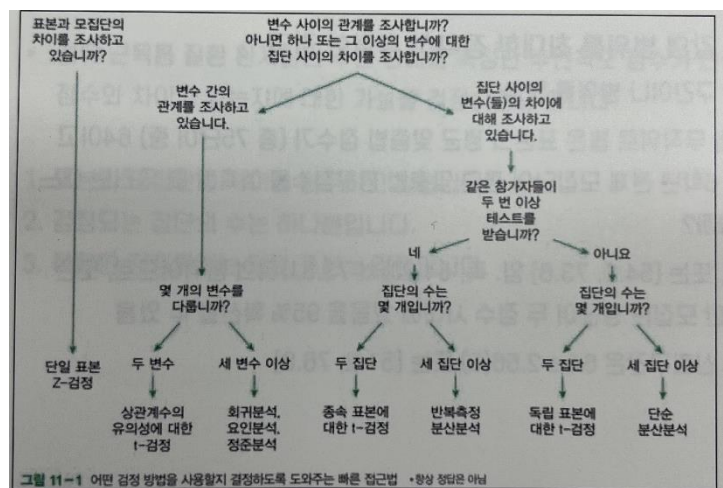
2) 데이터 수집

3) 유의확률과 유의수준 비교

4) 결론

### 2. 적합한 추론 통계 검정 방법 선택

(우측 순서도 참고)





-순서도 확인 순서

: 단일 표본인지 > 차이/관계 무엇을 조사하는지 > 독립인지 짝지은 표본인지

-우리가 배울 추론 통계 방법

: 단일표본 z검정 > 독립 표본에 대한 t검정 > 종속 표본에 대한 t검정 > 단순 분산분석 > 상관 계수의 유의성에 대한 t검정 > 회귀분석

### 3. 유의성 검정(≒가설 검정)

\*가설 검정의 순서 7개

1) 가설 설정 :  $H_0$ 과  $H_1$  설정

2) 유의수준 설정 :  $\alpha=0.05$  / 0.01 / 0.001 > 문제에서 주어지지 않는다면  $\alpha=0.05$

3) 적절한 가설 검정 방법 선택 : 앞 페이지 순서도 확인(교재 197p.)

4) 검정 통계량, 유의확률 계산 : 검정 통계량 공식 이용, 유의확률 계산

5) 임계값, 유의수준을 비교 기준 값으로 설정

-임계값은 검정통계량과 / 유의수준은 유의확률과 비교

-우리가 원하는 결과 값은 오른쪽에 있음 > 우측 검정만 확인하기

6) 검정통계량과 임계값, 유의확률과 유의수준 비교

-검정통계량과 임계값 : 축의 값 > 오른쪽으로 갈수록 커짐

-유의확률과 유의수준 : 면적(확률)

7) 결론

-임계값 < 검통량 = 유의수준 > 유의확률 : 기각역, 귀무가설 기각, 차이가 있다

-임계값 > 검통량 = 유의수준 < 유의확률 : 채택역, 귀무가설 채택, 차이가 없다

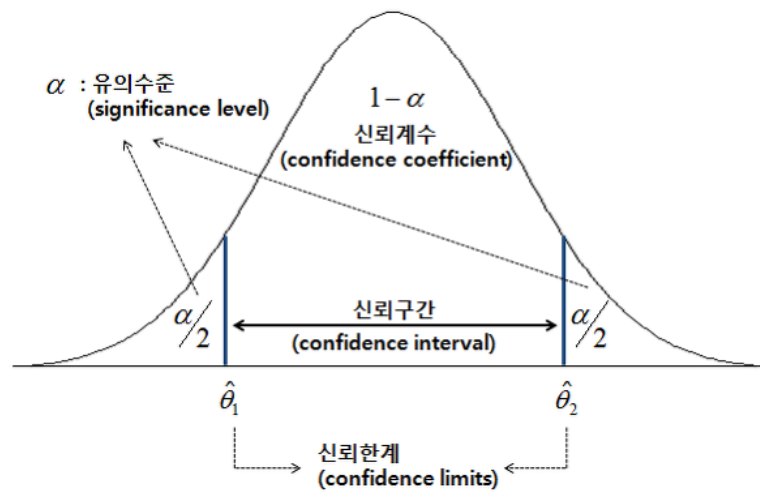
---

### 4. 주어진 표본의 값으로 모집단의 값의 범위를 최대한 정확히 추정한 결과

-가설검정 > 신뢰구간

-신뢰수준에 해당하는 축의 값의 구간이나 범위를 말함 (구간 : 신뢰구간)

[ 구간 추정 (Interval Estimation) ]



[R 분석과 프로그래밍] <http://rfriend.tistory.com>

5. 단일표본 z검정

<문제>

1.

(a) 정규 곡선의 3가지 특성

- 1) 평균=중앙값=최빈값
- 2) 평균을 중심으로 좌우 대칭
- 3) 꼬리가 점근적이다(점근적 꼬리) > 축에 닿지 않는다

(b)

인간의 어떤 행동, 특성 또는 성질이 정규분포되어 있다고 생각할 수 있습니까?

: O

2. 왜 z점수가 표준점수입니까? 어떻게 z점수를 사용하여 서로 다른 분포의 점수를 비교할 수 있습니까?

\*공식 = (원점수-평균)/표준편차

- 1) 동일한 단위를 사용하기 때문에

2) 평균, 표준편차를 이용해서 표준화하기 때문에(표준화하면 단위가 같아짐)

3. z점수를 사용하는 주요 이유는 무엇입니까? 어떻게 z점수를 사용할 수 있는지 예를 들어 보세요.

: 서로 다른 분포의 비교가 가능하기 때문

4. 평균은 50이고 표준편차가 5일 때, 다음 원점수에 대한 z점수를 계산하세요.

(a) 55    (b) 50    (c) 60    (d) 58.5    (e) 46

<엑셀 풀이>

<파이썬 풀이>

```
[16] # 4번
      x = pd.Series([55,50,60,58.5,46])

[17] x_bar = 50;x_sd=5

[18] # z점수
      (x - x_bar)/x_sd

0    1.0
1    0.0
2    2.0
3    1.7
4   -0.8
dtype: float64
```

5. 1.5의 z점수, 평균 40점, 표준편차 5점이 주어지면 해당 원점수는 얼마입니까?

-**관측치 = z score \* 표준편차 + 평균**

: 47.5점

<파이썬>

```
[20] # 5번
      # 원점수(관측치) = z점수*표준편차+평균
      x_bar = 40;x_sd = 5;z_score = 1.5

[21] z_score * x_sd + x_bar

47.5
```

6. 다음 질문은 평균 75, 표준편차 6.38을 갖고 있는 점수 분포를 전제로 합니다. 필요할 경우 간단한 그림으로 표현하세요.

(a) 점수가 70점에서 80점 사이에 있을 확률 :

(b) 점수가 80 점을 넘을 확률 :

(c) 점수가 81 점과 83 점 사이에 있을 확률 :

(d) 점수가 원점수 63 이하로 떨어질 확률 :

<엑셀 이용>

<파이썬 이용>

```
[22] # 6번
x = pd.Series([70,80,81,83,63])

[23] x_bar = 75;x_sd = 6.38

[29] # z점수
z_scoe = (x - x_bar)/x_sd

[31] # 누적확률 = 누적분포함수
# 왼쪽에서 오른쪽으로 가면서 누적
cum_p = norm.cdf(z_scoe);cum_p

array([0.21660836, 0.78339164, 0.82650375, 0.89506418, 0.02999428])

[33] tab = pd.DataFrame({'x':x,
                        'z_score':z_score,
                        'cum_p':cum_p})

[34] tab
```

	x	z_score	cum_p
0	70	1.5	0.216608
1	80	1.5	0.783392
2	81	1.5	0.826504
3	83	1.5	0.895064
4	63	1.5	0.029994

✓ 초	[39] # 1) 70과 80 사이 cum_p[1] - cum_p[0]	0.5667832855106969
✓ 초	[36] # 2) 80보다 높을 확률 1 - cum_p[2]	0.17349624562146615
✓ 초	[37] # 3) 81과 83 사이 cum_p[3] - cum_p[2]	0.06856042975493926
✓ 초	[38] # 4) 63 이하 cum_p[4]	0.029994275757256352

-엑셀 Z점수 함수(누적확률을 Z점수로) : NORM.S.INV

-유의성이 연구 및 추론통계의 활용에 중요한 구성 요소인 이유

: 유의성은 분석한 결과에서 유의한 차이가 있는지 판단할 수 있는 기준이 되기 때문

: 유의수준 > 유의확률 = 오른쪽에 있다 = 기각역 = 귀무가설을 기각한다 = 연구가설 채택 = 통계적으로 유의한 차이가 있다

-1종오류 = 유의수준 = 확률은 아무리 작아도 존재한다

-유의수준은 결과와 아무 관련이 없음 / 단지 연구하기 전에 설정하는 값에 불과함

-우연 = 유의확률

: 우연히 발생할 확률이 낮을수록 좋은 것

: 연구가설을 채택하려면 유의수준 > 유의확률