통계분석
(Statistics Analysis)

목차

1. 가설(Hypothesis)
2. 가설검정
3. 정규성검정
4. 이항검정
5. 카이제곱검정
6. T검정(단일표본,독립표본,대응표본)
7. 상관분석
8. 회귀분석

1. 가설(Hypothesis)

- **가설(假說, Hypothesis)?**
  - 이미 알려진 상황을 설명하기 위해서 설정한 <u>가정</u>
    - 과거부터 믿어 온 관습이나 관행

    예) 대한민국 수도는 서울이다.
  - 어떤 문제를 검증하기 위해서 <u>미리 세운 결론</u>
  - 주어진 연구 문제에 대한 <u>예측적</u> 해답(잠정적 진술)
  - 통계분석을 통해서 채택 또는 기각(통계적 가설검정)

[가설 예] 2021년도 고등학교 3학년 남학생의 키는 175cm이다.

가설검정 예

- **통계적 가설검정 예**
  - 표본에서 얻은 정보를 통해서 귀무가설과 대립가설 중 어떤 가설이 옳고, 그른지를 확률적으로 결정

귀무가설(H_0) 2021년도 고등학교 3학년 남학생의 키는 175.3cm이다.

↓

[표본 & 통계량] 주요 10개 도시를 대상으로 1,000명씩 표본으로 선정하여 평균 키를 계산한다.

↓

[가설검정] 가설을 지지하는 확률에 따라서 채택 or 기각

가설 유형

- 가설 유형

1. 귀무가설(영가설)

   ‘ 두 변수간의 관계가 없다.’ 또는 ‘차이가 없다.’(‘효과가 없다.’)

   - 부정적 형태 진술, 사실과 같다.
   - 예1) H_0 : 교육수준에 따라서 만족도에 차이가 없다.
   - 예2) H_0 : 2020년도 고3 남학생의 키는 175cm이다.

2. 대립가설(연구가설)

   ‘ 두 변수간의 관계가 있다.’, ‘차이가 있다.’(‘효과가 있다.’)

   - 긍정적 형태 진술, 사실과 다르다.
   - 예1) H_1: 교육수준에 따라서 만족도에 차이가 있다.
   - 예2) H_1 : 2020년도 고3 남학생의 키는 175cm가 아니다.

※ 귀무가설을 기준으로 가설검정을 수행한다.

가설 설정하기

- 귀무가설 설정 방법

[예] 법정 형사 재판에서 피고인에 대한 영가설은 다음 중 어느 것이 합리적일까?

~~H0 : 피고인은 유죄다.~~

H0 : 피고인은 무죄다.

정답) 2번의 가설을 반증하기 위해서 뚜렷한 증거를 제시할 수 있지만,

1번은 피고인의 무죄를 입증하기 어렵다.

❖ 증거재판주의 : 무죄의 가능성을 생각하기 어려울 정도의 엄격한 증명이 있어야 한다.

귀무가설(영가설)은 뚜렷한 증거를 제시하지 못하면 기각할 수 없다.

가설 검정방법

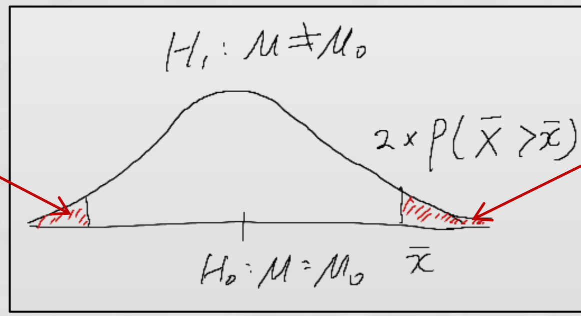
대립가설 기준으로 결정

- 양측검정(two-side test) : 대립가설(H_1)에 <u>방향성이 없는</u> 경우
  예) H_1: $\mu \neq$ 50kg
- 단측검정(one-side test) : 대립가설(H_1)에 <u>방향성이 포함</u>되는 경우
  예) H_1: μ > 50kg 또는 H_1: μ < 162cm

- 양측검정 : alternative='two-sided'
- 우측검정 : alternative= 'geater'
- 좌측검정 : alternative = 'less'

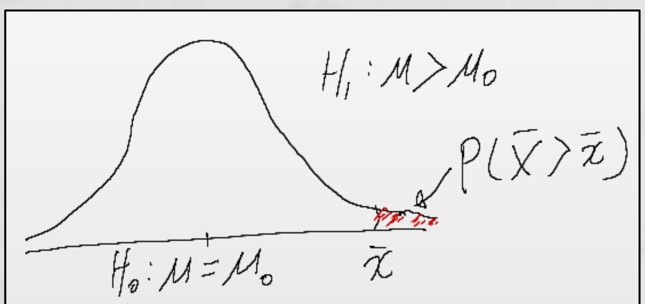
모평균 μ, 관측된 값 μ_0

기각역(H1 영역)

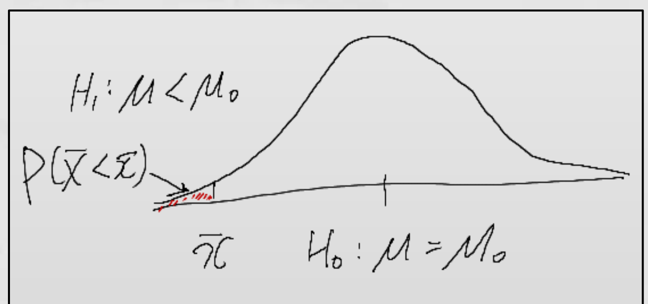


기각역(H1 영역)

우측검정



좌측검정



가설 검정방법

- 양측검정(two-side test) : 대립가설(H_1)에 방향성이 없는 경우

  H_0: 같다

  H_1: 다르다(같지 않다)

  - 양측검정 : alternative='two-sided'

- 우측검정(one-side test) : 대립가설(H_1)에 방향성이 포함되는 경우

  H_0: 같다

  H_1: 차이가 0보다 크다

  - 우측검정 : alternative= 'geater'

- 좌측검정(one-side test) : 대립가설(H_1)에 방향성이 포함되는 경우

  H_0: 같다

  H_1: 차이가 0보다 작다

  - 좌측검정 : alternative = 'less'

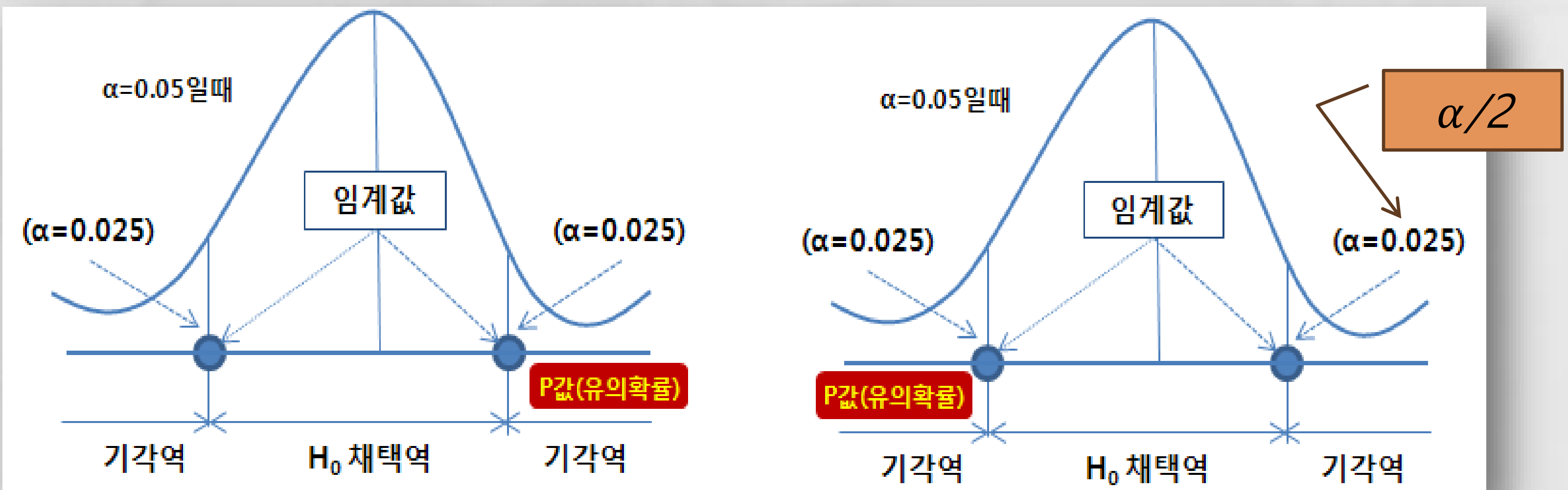
가설 검정방법

- 양측검정(2-sided test) : H_1에 방향성이 없는 가설 검정

H_0 : 성별에 따라 만족도에 차이가 없다.(남=여)

H_1 : 성별에 따라 만족도에 차이가 있다.(남≠여) ▶ 양측검정

❖ 방향성을 갖지 않은 대립가설 : ≠

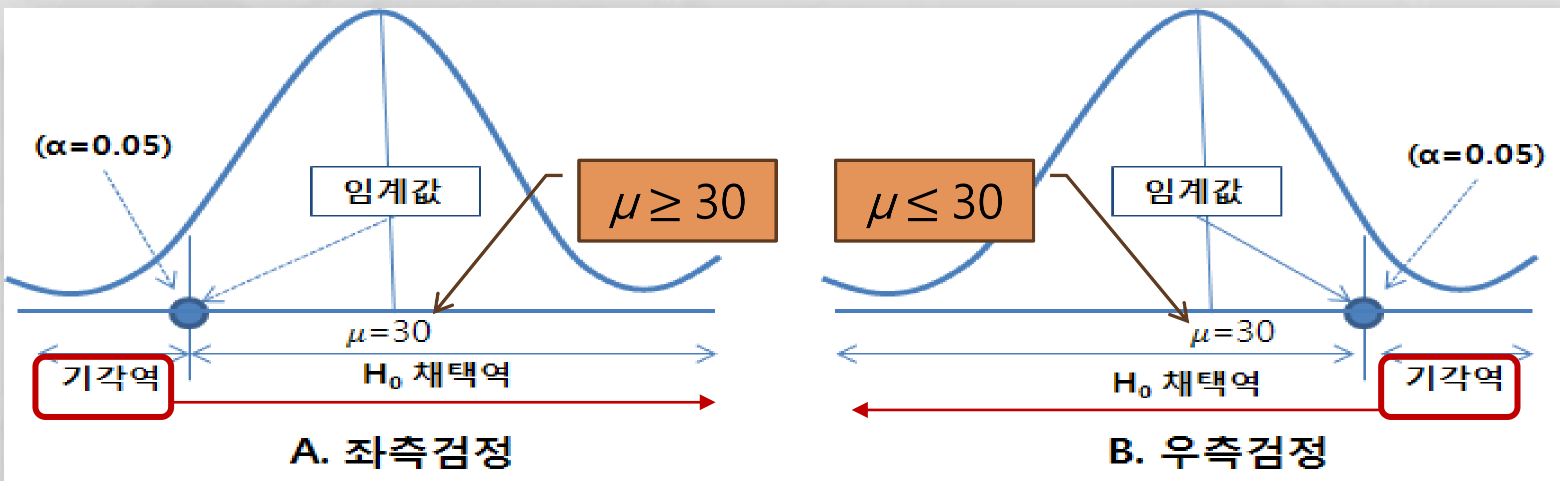


가설 검정방법

- 단측검정(1-sided test) : H_1에 방향성이 있는 가설 검정

H_0 : 1일 생산되는 불량품의 개수는 평균 30개 이다.(μ=30)
H_1 : 1일 생산되는 불량품의 개수는 평균 30개 이하이다.(μ<30) ▶ 좌측 단측검정
H_1 : 1일 생산되는 불량품의 개수는 평균 30개 이상이다.(μ>30) ▶ 우측 단측검정

- 방향성을 갖는 두 가지 대립가설 : < 또는 >



왼쪽 단측검정

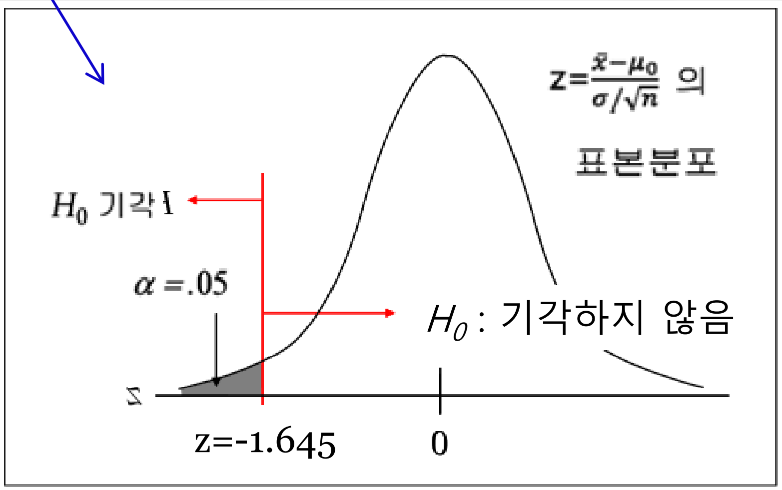
오른쪽 단측검정

가설 검정방법

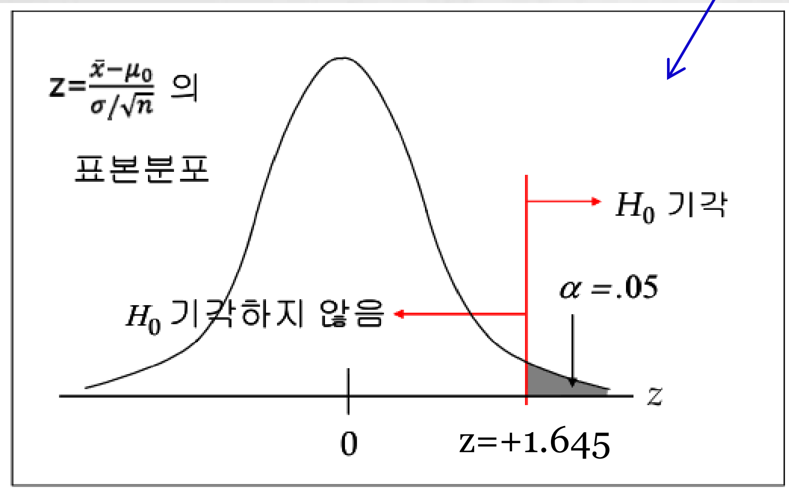
- 단측검정 : 임계값 1개

표준정규분포에서 95% 신뢰수준 :
임계값 : α = 5%(0.05)
채택역 : z값 > -1.645

표준정규분포에서 95% 신뢰수준 :
임계값 : α = 5%(0.05)
채택역 : z값 < 1.654



좌측검정



우측검정

95% 신뢰수준 일 때 α=5%에 해당하는 z값은 ±1.645 이다.

가설 설정 규칙

- 가설의 부등호 규칙

✓ 귀무가설(영가설)은 등호가 반드시 포함되어야 하고, 대립가설은 절대로 등호가 포함되지 않아야 한다. 또한 표본의 검정통계량이 아닌 모집단의 모수로 표현한다.

예1) 대립가설 H_1: $\mu < 5$kg 일 때 귀무가설 H_0: $\mu \geq 5$kg 또는 H_0: $\mu = 5$kg 가능

예2) 대립가설 H_1: $\mu > 161$cm 일 때 귀무가설 H_0: $\mu \leq 161$cm 또는 H_0: $\mu = 161$cm

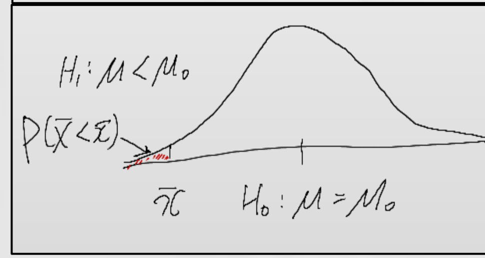
가설 설정 사례

[예1] 한 헬스 클럽이 3개월 안에 평균 체중을 <u>5kg 이상 </u>줄일 수 있다고 광고합니다. 이 클럽 회원 45명의 체중 감량을 조사하였더니 평균 4.8kg입니다. 모든 헬스 클럽 이용자들의 평균 체중 감량을 μ 라고 표현할 때 설정한 가설은?

H0 : μ >= 5 or μ = 5
H1 : μ < 5 → **단측검정**

해설) 평균 체중 5kg 이상은 기존에 알려진 사실

모평균: μ, 실제 값: μ_0

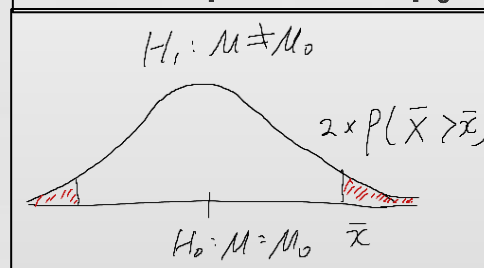


[예2] 2021년 정당 A의 지지율은 32% 입니다. 2022년 지지율이 달라졌는지 알아보기 위하여 여론조사를 합니다. 2021년 정당 A의 지지율을 p라고 표기할 때 가설은?

H0 : p=0.32%
H1 : p ≠ 0.32% → **양측검정**

해설) 지지율 32%는 기존에 알려진 사실

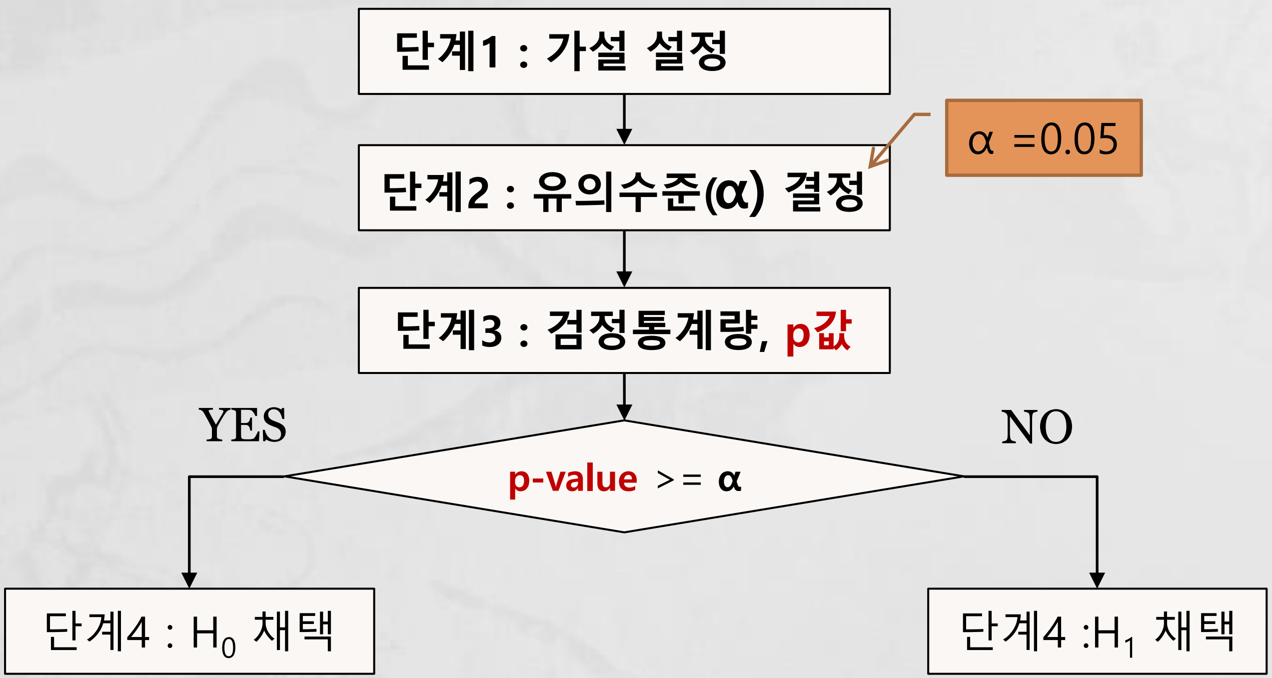
모평균: μ, 실제 값: μ_0



2. 가설검정

✓ 유의확률(p-value)이 유의수준(α) 보다 크면 가설이 채택되고, 유의수준 보다 적으면 가설이 기각(통계적으로 유의하다.)된다.

● **가설검정 절차**



가설검정 사례

- 두 집단 평균차이 검정 : t검정(양측검정) 이용

주어진 데이터에는 여학생과 남학생 각각 30명씩 국어점수를 저장하고 있다. 두 집단 간 평균에 차이가 있는지 답하시오. 가설은 아래와 같다.

귀무가설(H_O) : 여학생 점수평균 = 남학생 점수 평균

대립가설(H_1) :여학생 점수평균 ≠ 남학생 점수 평균

```
female_score = np.random.uniform(50, 100, size=30) # 여학생 점수
male_score = np.random.uniform(45, 95, size=30) # 남학생 점수

two_sample = stats.ttest_ind(female_score, male_score,
                                        alternative='two-sided')
print(two_sample)
print( ' 검정통계량 = %.3f, p-value = %.3f'%(two_sample))
# 검정통계량 = 0.159, p-value = 0.874

# 가설검정 : 유의수준 5% 수준에서 남.녀 학생의 평균 점수에 차이가 없다.(채택)
```

가설검정 사례

- 두 집단 평균차이 검정 : t검정(단측검정) 이용

주어진 데이터에는 고혈압 환자 120명의 치료 전후의 혈압이 저장되어 있다. 해당 치료가 효과가 있는지(즉 치료 후의 혈압이 감소했는지) 대응표본 t-검정 (paired t-test)를 통해 답하고자 한다. 가설은 아래와 같다.

μd : (치료 후 혈압 - 치료전 혈압)의 평균

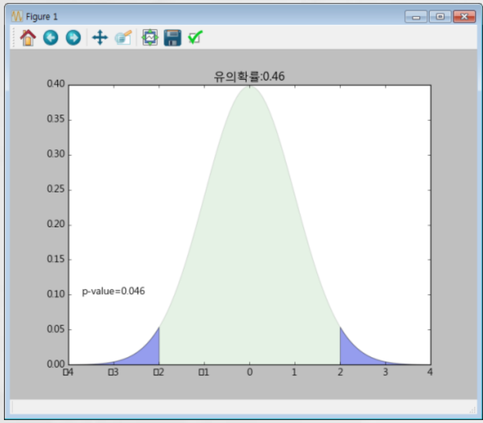
H_0 : μd >= 0

H_1 : μd < 0 : 방향성을 갖는 대립가설(0보다 작다)

```
from scipy import stats # 가설검정

result = stats.ttest_rel(치료전혈압, 치료후혈압, alternative='less')
result
'''
statistic=3.3371870510833657, pvalue=0.0005648957322420411)
'''
# 가설검정 : 유의수준 5% 수준에서 혈압이 감소했다고 할 수 있다.(기각)
```

3. 정규성 검정(Normality Test)



데이터의 분포가 정규분포(Normal Dstribution)를 따르는지 검정

```
# 1) 표준정규분포 생성
mu, sigma = 0, 1
norm_obj = stats.norm(mu, sigma)
print(norm_obj) # objec info

# 2) 확률변수 X : 시행횟수 N번으로 정규분포의 확률변수 만들기
N = 1000 # sample 수
X = norm_obj.rvs(size = N) # rvs(random variable sampling) : N번 시뮬레이션

# 3) 정규성 검정
# 귀무가설(H0) : 정규분포와 차이가 없다.
print(stats.shapiro(X))
'''
 statistic=0.9983075261116028
 pvalue=0.4358219504356384)
'''
```

가설검정 : 유의수준 5% 수준에서 정규분포와 차이가 없다.(채택)

4. 이항검정(binominal test)

- ❖ 이항검정(binominal test) : 이항분포를 이용한 가설검정으로, 이항분포는 2가지 범주(성공/실패)를 갖는 이산확률분포이다.
- ❖ 베르누이 확률 분포 : 이항분포에서 '성공' 확률을 모수로 갖는 확률분포
- ❖ 이항분포 : 베르누이 시행을 적용한 확률분포를 말한다.
- ▪ 베르누이 분포 : B(N=1, P) -> 독립시행(확률실험) 1회
- ▪ 이항분포 : B(N=n, P) -> 베르누이 독립시행 n번

<연구환경>

150명의 합격자 중에서 남자 합격자가 62명일 때 99% 신뢰수준에서 남.여 합격률에 차이가 있다고 할수 있는가?

H_0 : 남여 합격률에 차이가 없다.(p=0.5)

H_1 : 남여 합격률에 차이가 있다.

성공횟수 시행횟수 모수확률 양측검정

```
x = 62 # 성공회수
pvalue = stats.binom_test(x=x, n=150, p=0.5, alternative='two-sided')
print('n = %d, pvalue= %.5f'%(x, pvalue))
# n = 62, pvalue= 0.04087
```

5. 카이제곱검정(chisqure test)

- ❖ 범주(Category)별로 관측 빈도와 기대빈도가 차이가 있는지 검정
- ❖ 카이제곱 분포에 기초한 통계적 방법(카이제곱 분포표 이용)
- ❖ χ_2 = Σ (관측값 - 기댓값)2 / 기댓값
- ❖ - 분석을 위해서 교차분할표 작성
- ❖ 교차분석은 검정통계량으로 카이제곱 사용(=카이제곱 검정)
- ❖ 검증 유형 분류 : 일원카이제곱검정, 이원카이제곱검정

① 일원카이제곱 검정 : 1개 변수 이용

귀무가설 : 관측치와 기대치는 차이가 없다.(게임에 적합하다.)
대립가설 : 관측치와 기대치는 차이가 있다.(게임에 적합하지 않다.)

```
real_data = [4, 6, 17, 16, 8, 9] # 관측치
exp_data = [10,10,10,10,10,10] # 기대치
chis = stats.chisquare(real_data, exp_data)
print('statistic = %.3f, pvalue = %.3f'%(chis))
# statistic = 14.200, pvalue = 0.014
```

statistic -> $\chi^2 = \Sigma$ (관측값 - 기댓값)2 / 기댓값

② 이원카이제곱 검정 : 2개 변수 이용

교차분할표의 관측값과 χ^2 계산식에 의해서 구한 기댓값으로 검정 수행

χ^2 = Σ (관측값 - 기댓값)2 / 기댓값

귀무가설 : 교육수준과 흡연율 간에 관련성이 없다.(채택)

대립가설 : 교육수준과 흡연율 간에 관련성이 있다.(기각)

x \ y	1	2	3	Row Total	
1	51	92	68	211	관측값
	68.94	83.80	58.24		기대값=행합계 * 열합계 / 총합계
2	22	21	9	52	관측값
	16.99	20.65	14.35		기대값
3	43	28	21	92	관측값
	30.06	36.54	25.39		기대값
Column Total	116	141	98	355	

```
# <단계 1> 변수 선택
print(smoke)# education, smoking 변수
education = smoke.education # smoke['education']
smoking = smoke.smoking # smoke['smoking']

# <단계 2> 교차분할표
tab = pd.crosstab(index=education, columns=smoking)
print(tab) # 관측값
'''
smoking    1   2   3
education
1         51  92  68
2         22  21   9
3         43  28  21
'''
# <단계3> 카이제곱 검정 : 교차분할표 이용
chi2, pvalue, df, evalue = stats.chi2_contingency(observed= tab)
# chi2 검정통계량, 유의확률, 자유도, 기대값
print('chi2 = %.6f, pvalue = %.6f, d.f = %d'%(chi2, pvalue, df))
# chi2 = 18.910916, pvalue = 0.000818, d.f = 4
```

[해설] 유의미한 수준에서 교육수준과 흡연율 간에 관련성이 있다고 볼 수 있다.
(기대치와 관찰치는 차이가 있다.)

6 T-검정

- **Z-검정**
  - ✓ 모집단 정규분포이고, 모집단의 분산(표준편차)이 알려진 경우
  - ✓ 모집단의 표준편차를 이용하여 모평균 추정/검정(Z분포)
  - ✓ 기본 가정 : 정규분포
  - ✓ 기본 가설 : 모평균과 차이가 없다.

Z통계량 〈정규분포〉

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- **T-검정**
  - ✓ **모집단 정규분포이고, 모집단의 분산(표준편차)이 알려지지 않은 경우**
  - ✓ 표본의 표준편차 이용하여 모평균 추정/검정(T분포)
  - ✓ 기본 가정 : 정규분포
  - ✓ 기본 가설 : 모평균과 차이가 없다.

T통계량 〈t분포〉

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

2. 모평균 검정

- 모평균 검정
  - ✓ 표본으로 모집단의 모평균을 추정/검정하는 방법

모집단 수	검정 대상	검정방법
1개	모집단의 모평균 추정 (표본 vs 모평균 차이 검정)	Z검정(모분산 알고 있는 경우)
		T검정(모분산 모르는 경우)
2개	두 집단의 평균차이 검정	Z검정(모분산 알고 있는 경우)
		T검정(모분산 모르는 경우)
3개 이상	세 집단 이상 평균차이 검정	분산분석(ANOVA)

6-1. 단일표본 t-검정

❖ 단일 표본 t-검정은 정규 분포의 표본에 대해 기댓값을 조사하는 검정방법

❖ 수집된 표본이 모평균과 차이가 있는지 검정하는 방법

❖ scipy stats 서브패키지의 ttest_1samp 함수를 사용

예) 평균 키가 176이라고 할 수 있는지 유의수준 5%로 검정

```
# 단일집단 평균차이 검정
one_group_test = stats.ttest_1samp(sample_data, 176)
print('t검정 통계량 = %.3f, pvalue = %.5f'%(one_group_test))
# t검정 통계량 = 1.255, pvalue = 0.21972
```

6-2. 독립표본 t-검정

❖ 독립 표본 t-검정(Independent-two-sample t-test)은 간단하게 two sample t-검정

❖ 두 개의 독립적인 정규 분포에서 나온 두 개의 데이터 셋을 사용하여 두 정규 분포의 기댓값이 동일한지를 검사

❖ scipy stats 서브패키지의 ttest_ind 함수를 사용

❖ 독립 표본 t-검정은 두 정규 분포의 분산 값이 같은 경우와 같지 않은 경우에 사용하는 검정 통계량이 다르기 때문에 equal_var 인수를 사용하여 이를 지정

❖ 서로 다른 10명의 사람에게 수면제1을 복용했을 때의 수면 증가 시간은 [0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0] 이고 수면제2를 복용했을 때의 수면 증가 시간은 [1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4] 인 경우 2가지 약 복용 시 수면 증가 시간은 차이가 없는지 유의확률 5%로 검정

```
import numpy as np
from scipy import stats
import scipy as sp

x1 = np.array([0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0]);
x2 = np.array([1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4]);
r = sp.stats.ttest_ind(x1, x2, equal_var=True)
print(x1.var())

if r.pvalue >= 0.05:
    print("2가지 약의 평균 수면 증가시간은 같다.")
else:
    print("2가지 약의 평균 수면 증가시간은 다르다.")
```

6-3. 대응표본 t-검정

- ❖ 대응 표본 t-검정은 독립 표본 t-검정을 두 집단의 샘플이 1대1 대응하는 경우에 대해 수정한 것
- ❖ 독립 표본 t-검정과 마찬가지로 두 정규 분포의 기댓값이 같은지 확인하기 위한 검정
- ❖ 예를 들어 어떤 반의 학생들이 특강을 수강하기 전과 수강한 이후에 각각 시험을 본 시험 점수의 경우에는 같은 학생의 두 점수는 대응
- ❖ 이 대응 정보를 알고 있다면 보통의 독립 표본 t-검정에서 발생할 수 있는 샘플간의 차이의 영향을 없앨 수 있기 때문에 특강 수강의 영향을 보다 정확하게 추정
- ❖ scipy stats 서브패키지의 ttest_rel 함수를 사용

```
import numpy as np
from scipy import stats
import scipy as sp

x1 = np.array([0.7,-1.6,-0.2,-1.2,-0.1,3.4,3.7,0.8,0.0,2.0]);
x2 = np.array([1.9,0.8,1.1,0.1,-0.1,4.4,5.5,1.6,4.6,3.4]);
r = sp.stats.ttest_rel(x1, x2)
print(x1.var())

if r.pvalue >= 0.05:
    print("2가지 약의 평균 수면 증가시간은 같다.")
else:
    print("2가지 약의 평균 수면 증가시간은 다르다.")
```

7. 상관관계

- ❖ 두 변수 간에 어떤 선형적 관계가 있는지 분석하는 것을 이를 상관 분석 (Correlation Analysis)이라고 한다.
- ❖ 예를 들면 교육수준과 월 수입 간의 관계
- ❖ 공분산은 결합 분포의 평균을 중심으로 각 자료들이 어떻게 분포되어 있는지를 보여준다.

샘플 공분산(sample covariance)은 다음과 같이 정의된다. 여기에서 x_i와 y_i는 각각 i번째의 x 자료와 y자료의 값을 가리키고, m_x와 m_y는 x 자료와 y자료의 샘플 평균을 가리킨다.

$$s_{xy}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)$$

- 분포의 크기는 공분산이 아닌 분산만으로도 알 수 있기 때문에 대부분의 경우 자료 분포의 방향성만 분리하여 보는 것이 유용한데 이 때 필요한 것이 상관계수 (correlation coefficient)
- 값의 범위는 -1에서 1사이입니다.
- 두 변수 사이에는 강한 양의 상관관계가 존재하면 상관 계수의 값은 1에 가까울 것입니다. 반대로 두 변수 사이에 특별한 상관관계가 존재하지 않으면 상관 계수의 값은 0에 가까울 것이며, 두 변수 사이에 음의 상관관계가 존재한다면 상관 계수의 값은 -1에 가까워진다.
- DataFrame 이나 Series 객체의 corr()을 이용하면 상관계수를 알아볼 수 있다.
- cov()는 공분산

```
From pandas import Series, DataFrame

sales = Series([3,5,8,11,13])
dms = Series([1,2,3,4,5])
su = sales.corr(dms)
print("상관계수:",su)
```

8. 회귀분석

❖ 두 변수 사이에 상관관계가 존재한다면 통계적으로 모델을 작성할 수 있으며, 특히 단순하게 일차원적인 선형 모델의 경우라면 다음과 같이 단순한 수식 형태로 표현할 수 있다

Yi = B0 + B1 * Xi + Ei

❖ 여기서 Yi를 종속 변수(dependent variable), Xi를 독립 변수(independent variable)라고 합니다. 종속 변수와 독립 변수 간의 관계식을 결정하는 B0와 B1는 데이터로부터 추정할 수 있으며, 이를 각각 절편(intercept), 기울기(slope)라고 합니다. Ei는 오차항으로 우리가 작성한 모델과 실제 데이터 값과의 차이를 나타낸다

- ❖ 종속 변수에 영향을 주는 변수가 1개일 경우를 단일 회귀분석
- ❖ 단일 회귀 분석의 경우는 scipy 패키지의 stats 모듈의 linregress 함수 이용
- ❖ 결과는 float 형으로 선형 모델의 기울기, 절편, 상관 계수, p-value, 에러의 표준 편차가 순차적으로 반환
- ❖ 여기서 p-value는 통계학에서 예측 불확실성의 정도를 나타내는 값으로, 일반적으로 0.05 미만일 때가 통계학적으로 유의미 함

```
x = data['production']
y = data['quantity'])
slope, intercept, r_value, p_value, stderr = stats.linregress(x,  y)
print("기울기:", slope)
print("절편:", intercept)
print("상관계수", r_value)
print("불확실성 정도:", p_value)
print("생산금액이 4가 되기 위한 전기 사용량:", end=' ')
print(4 * slope + intercept)
```

다중회귀분석

- ❖ statsmodels 는 통계 분석을 위한 python 패키지
- ❖ http://www.statsmodels.org
- ❖ statsmodels는 기초 통계, 회귀 분석, 시계열 분석 등 다양한 통계 분석 제공
- ❖ 기초 통계 (Statistics)
  - ✓ 각종 검정(test) 기능
  - ✓ 커널 밀도 추정
  - ✓ Generalized Method of Moments
- ❖ 회귀 분석 (Linear Regression)
  - ✓ 선형 모형 (Linear Model)
  - ✓ 일반화 선형 모형 (Generalized Linear Model)
  - ✓ 강인 선형 모형 (Robust Linear Model)
  - ✓ 선형 혼합 효과 모형 (Linear Mixed Effects Model)
  - ✓ ANOVA (Analysis of Variance)
  - ✓ Discrete Dependent Variable (Logistic Regression 포함)
  - ✓ 시계열 분석 (Time Series Analysis)

❖ 단일 회귀 분석

- ✓ statsmodels.regression.linear_model.OLS(endog, exog=None)
- ✓ 파라미터
  - endog : 종속 변수. 1차원 배열
  - exog : 독립 변수, 2차원 배열.
- ✓ statsmodels 의 OLS 클래스는 자동으로 상수 항을 만들어주지 않기 때문에 사용자가 add_constant 명령으로 상수 항을 추가해야 한다.
- ✓ 모형 객체가 생성되면 fit, predict 메서드를 사용하여 추정 및 예측을 실시합니다.
- ✓ 예측 결과는 RegressionResults 클래스 객체로 출력되며 summary 메서드로 결과 보고서를 볼 수 있다.

❖ 다중 회귀 분석

- ✓ statsmodels.formula.api 패키지의 ols(formula = '종속변수 ~ 독립변수[+ 독립변수] ' , data = 데이터프레임).fit()을 호출해서 결과를 리턴 받음
- ✓ 결과의 params 가 y절편과 각 독립변수 와 의 상관계수를 Series로 리턴
- ✓ 결과의 pvalues 가 유의 확률을 Series로 리턴
- ✓ 결과의 predict()이 예측 값을 ndarray 타입으로 리턴
- ✓ 결과의 rsquared가 반응 변수 변동의 백분율을 리턴하는데 일반적으로 값이 클수록 모형이 데이터를 더 잘 적합시킨다.
  - 항상 0%에서 100% 사이
  - R-제곱은 다중 회귀 분석에서 결정 계수 또는 다중 결정 계수

name	score	iq	academy	game	tv
A	90	140	2	1	0
B	75	125	1	3	3
C	77	120	1	0	4
D	83	135	2	3	2
E	65	105	0	4	4
F	80	123	3	1	1
G	83	132	3	4	1
H	70	115	1	1	3
I	87	128	4	0	0
J	79	131	2	2	3

```
from pandas import Series, DataFrame
from scipy import stats
import statsmodels.formula.api as sm
df = pd.read_csv("score.csv", encoding="ms949")
result = sm.ols(formula = 'score ~ iq + academy + game + tv', data = df).fit()
print('회귀계수 : ', result.params)
print('Pvalue :', result.pvalues)
print('Rsquared : ', result.rsquared)
```

```
회귀 계수 :
Intercept    24.722251
iq            0.374196
academy       3.208802
tv            0.192573

Pvalue :
Intercept    7.873829e-20
iq           1.459524e-41
academy      5.259783e-15
tv           5.259387e-01

'Rsquared : 0.9464476338905841
```

최소자승법 OLS Regression Results

Dep. Variable:	score	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	860.1
Date:	Mon, 19 Jun 2023	Prob (F-statistic):	1.50e-92
Time:	17:45:03	Log-Likelihood:	-274.84
No. Observations:	150	AIC:	557.7
Df Residuals:	146	BIC:	569.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.7223	2.332	10.602	0.000	20.114	29.331
iq	0.3742	0.020	19.109	0.000	0.335	0.413
academy	3.2088	0.367	8.733	0.000	2.483	3.935
tv	0.1926	0.303	0.636	0.526	-0.406	0.791

Omnibus:	36.802	Durbin-Watson:	1.905
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57.833
Skew:	1.252	Prob(JB):	2.77e-13
Kurtosis:	4.728	Cond. No.	2.32e+03